Advanced Analytical Models

Over 800 Models and 300 Applications from the Basel II Accord to Wall Street and Beyond

JOHNATHAN MUN

WILEY
John Wiley & Sons, Inc.

Advanced Analytical Models

Founded in 1807, John Wiley & Sons is the oldest independent publishing company in the United States. With offices in North America, Europe, Australia, and Asia, Wiley is globally committed to developing and marketing print and electronic products and services for our customers' professional and personal knowledge and understanding.

The Wiley Finance series contains books written specifically for finance and investment professionals as well as sophisticated individual investors and their financial advisors. Book topics range from portfolio management to e-commerce, risk management, financial engineering, valuation, and financial instrument analysis, as well as much more.

For a list of available titles, please visit our Web site at www.WileyFinance.com.

Advanced Analytical Models

Over 800 Models and 300 Applications from the Basel II Accord to Wall Street and Beyond

JOHNATHAN MUN

WILEY
John Wiley & Sons, Inc.

Copyright © 2008 by Johnathan Mun. All rights reserved.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 646-8600, or on the Web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at http://www.wiley.com/go/permissions.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

Designations used by companies to distinguish their products are often claimed as trademarks. In all instances where John Wiley & Sons, Inc., is aware of a claim, the product names appear in initial capital or all capital letters. Readers, however, should contact the appropriate companies for more complete information regarding trademarks and registration.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our Web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data:

Mun, Johnathan.
Advanced analytical models : over 800 models and 300 applications from the Basel II Accord to Wall Street and beyond / Johnathan Mun.
p. cm. — (Wiley finance series)
Includes index.
ISBN 978-0-470-17921-5 (cloth/dvd)
1. Finance—Mathematical models. 2. Risk assessment—Mathematical models. 3. Mathematical models. 4. Computer simulation. I. Title.
HG106.M86 2008
003′.3—dc22
2007039385

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

Dedicated to my wife, Penny. Without your encouragement, advice, and support, this modeling book would never have taken off.

"Delight yourself in the Lord and He will give you the desires of your heart."
—Psalms 37:4 (NIV)

Contents

Preface xv

Software Applications xvii

PART 1

Modeling Toolkit and Risk Simulator Applications 1

Introduction to the Modeling Toolkit Software 1
Introduction to Risk Simulator 2
Running a Monte Carlo Simulation 6
Using Forecast Charts and Confidence Intervals 16
Correlations and Precision Control 18
Tornado and Sensitivity Tools in Simulation 22
Sensitivity Analysis 29
Distributional Fitting: Single Variable and Multiple Variables 33
Bootstrap Simulation 35
Hypothesis Testing 40
Data Extraction, Saving Simulation Results, and Generating Reports 42
Regression and Forecasting Diagnostic Tool 42
Statistical Analysis Tool 52
Distributional Analysis Tool 54
Portfolio Optimization 56
Optimization with Discrete Integer Variables 70
Forecasting 72

1. Analytics—Central Limit Theorem 79
2. Analytics—Central Limit Theorem—Winning Lottery Numbers 84
3. Analytics—Flaw of Averages 88
4. Analytics—Mathematical Integration Approximation Model 93
5. Analytics—Projectile Motion 96
6. Analytics—Regression Diagnostics 100
7. Analytics—Ships in the Night 109
8. Analytics—Statistical Analysis 111
9. Analytics—Weighting of Ratios 123

10. Credit Analysis—Credit Premium 125
11. Credit Analysis—Credit Default Swaps and Credit Spread Options 127
12. Credit Analysis—Credit Risk Analysis and Effects on Prices 129
13. Credit Analysis—External Debt Ratings and Spread 131
14. Credit Analysis—Internal Credit Risk Rating Model 133
15. Credit Analysis—Profit Cost Analysis of New Credit 135

16. Debt Analysis—Asset-Equity Parity Model 137
17. Debt Analysis—Cox Model on Price and Yield of Risky Debt with Mean-Reverting Rates 138
18. Debt Analysis—Debt Repayment and Amortization 141
19. Debt Analysis—Debt Sensitivity Models 145
20. Debt Analysis—Merton Price of Risky Debt with Stochastic Asset and Interest 147
21. Debt Analysis—Vasicek Debt Option Valuation 149
22. Debt Analysis—Vasicek Price and Yield of Risky Debt 151

23. Decision Analysis—Decision Tree Basics 153
24. Decision Analysis—Decision Tree with EVPI, Minimax, and Bayes' Theorem 158
25. Decision Analysis—Economic Order Quantity and Inventory Reorder Point 169
26. Decision Analysis—Economic Order Quantity and Optimal Manufacturing 170
27. Decision Analysis—Expected Utility Analysis 172
28. Decision Analysis—Inventory Control 174
29. Decision Analysis—Queuing Models 176

30. Exotic Options—Accruals on Basket of Assets 178
31. Exotic Options—American, Bermudan, and European Options with Sensitivities 180
32. Exotic Options—American Call Option on Foreign Exchange 182
33. Exotic Options—American Call Options on Index Futures 184
34. Exotic Options—American Call Option with Dividends 186
35. Exotic Options—Asian Lookback Options Using Arithmetic Averages 188
36. Exotic Options—Asian Lookback Options Using Geometric Averages 189
37. Exotic Options—Asset or Nothing Options 190
38. Exotic Options—Barrier Options 191
39. Exotic Options—Binary Digital Options 193
40. Exotic Options—Cash or Nothing Options 195
41. Exotic Options—Chooser Option (Simple Chooser) 196
42. Exotic Options—Chooser Option (Complex Chooser) 197
43. Exotic Options—Commodity Options 198
44. Exotic Options—Currency (Foreign Exchange) Options 199

45. Exotic Options—Double Barrier Options 200
46. Exotic Options—European Call Option with Dividends 201
47. Exotic Options—Exchange Assets Option 203
48. Exotic Options—Extreme Spreads Option 204
49. Exotic Options—Foreign Equity–Linked Foreign Exchange Options in Domestic Currency 205
50. Exotic Options—Foreign Equity Struck in Domestic Currency 207
51. Exotic Options—Foreign Equity with Fixed Exchange Rate 208
52. Exotic Options—Foreign Takeover Options 209
53. Exotic Options—Forward Start Options 210
54. Exotic Options—Futures and Forward Options 211
55. Exotic Options—Gap Options 212
56. Exotic Options—Graduated Barrier Options 213
57. Exotic Options—Index Options 214
58. Exotic Options—Inverse Gamma Out-of-the-Money Options 215
59. Exotic Options—Jump-Diffusion Options 216
60. Exotic Options—Leptokurtic and Skewed Options 217
61. Exotic Options—Lookback with Fixed Strike (Partial Time) 218
62. Exotic Options—Lookback with Fixed Strike 219
63. Exotic Options—Lookback with Floating Strike (Partial Time) 220
64. Exotic Options—Lookback with Floating Strike 221
65. Exotic Options—Min and Max of Two Assets 222
66. Exotic Options—Options on Options 223
67. Exotic Options—Option Collar 224
68. Exotic Options—Perpetual Options 225
69. Exotic Options—Range Accruals (Fairway Options) 226
70. Exotic Options—Simple Chooser 228
71. Exotic Options—Spread on Futures 229
72. Exotic Options—Supershare Options 230
73. Exotic Options—Time Switch Options 231
74. Exotic Options—Trading-Day Corrections 232
75. Exotic Options—Two-Asset Barrier Options 233
76. Exotic Options—Two Asset Cash or Nothing 234
77. Exotic Options—Two Correlated Assets Option 235
78. Exotic Options—Uneven Dividend Payments Option 237
79. Exotic Options—Writer Extendible Option 238

80. Forecasting—Data Diagnostics 239
81. Forecasting—Econometric, Correlations, and Multiple Regression Modeling 248
82. Forecasting—Exponential J-Growth Curves 254
83. Forecasting—Forecasting Manual Computations 257
84. Forecasting—Linear Interpolation and Nonlinear Spline Extrapolation 259
85. Forecasting—Logistic S-Growth Curves 264

86. Forecasting—Markov Chains and Market Share 267
87. Forecasting—Multiple Regression 269
88. Forecasting—Nonlinear Extrapolation and Forecasting 271
89. Forecasting—Stochastic Processes, Brownian Motion, Forecast Distribution at Horizon, Jump-Diffusion, and Mean-Reversion 273
90. Forecasting—Time-Series ARIMA 276
91. Forecasting—Time-Series Analysis 283

92. Industry Applications—Biotech Manufacturing Strategy 287
93. Industry Applications—Biotech Inlicensing Drug Deal Structuring 289
94. Industry Applications—Biotech Investment Valuation 291
95. Industry Application—Banking: Integrated Risk Management, Probability of Default, Economic Capital, Value at Risk, and Optimal Bank Portfolios 293
96. Industry Application—Electric/Utility: Optimal Power Contract Portfolios 321
97. Industry Application—IT—Information Security Intrusion Risk Management 329
98. Industry Applications—Insurance ALM Model 349

99. Operational Risk—Queuing Models at Bank Branches 354

100. Optimization—Continuous Portfolio Allocation 356
101. Optimization—Discrete Project Selection 362
102. Optimization—Inventory Optimization 366
103. Optimization—Investment Portfolio Allocation 372
104. Optimization—Investment Capital Allocation I (Basic Model) 373
105. Optimization—Investment Capital Allocation II (Advanced Model) 376
106. Optimization—Military Portfolio and Efficient Frontier 380
107. Optimization—Optimal Pricing with Elasticity 386
108. Optimization—Optimization of a Harvest Model 390
109. Optimization—Optimizing Ordinary Least Squares 394
110. Optimization—Stochastic Portfolio Allocation 400

111. Options Analysis—Binary Digital Instruments 405
112. Options Analysis—Inverse Floater Bond 407
113. Options Analysis—Options-Trading Strategies 413
114. Options Analysis—Options-Adjusted Spreads Lattice 420
115. Options Analysis—Options on Debt 422
116. Options Analysis—Five Plain Vanilla Options 424

117. Probability of Default—Bond Yields and Spreads (Market Comparable) 432
118. Probability of Default—Empirical Model 434
119. Probability of Default—External Options Model (Public Company) 437

120. Probability of Default—Merton Internal Options Model (Private Company) 441
121. Probability of Default—Merton Market Options Model (Industry Comparable) 442

122. Project Management—Cost Estimation Model 443
123. Project Management—Critical Path Analysis (CPM PERT GANTT) 446
124. Project Management—Project Timing 453

125. Real Estate—Commercial Real Estate ROI 456

126. Risk Analysis—Integrated Risk Analysis 460
127. Risk Analysis—Interest Rate Risk 472
128. Risk Analysis—Portfolio Risk Return Profiles 474

129. Risk Hedging—Delta-Gamma Hedging 477
130. Risk Hedging—Delta Hedging 478
131. Risk Hedging—Effects of Fixed versus Floating Rates 479
132. Risk Hedging—Foreign Exchange Cash Flow Model 481
133. Risk Hedging—Hedging Foreign Exchange Exposure 487

134. Sensitivity—Greeks 491
135. Sensitivity—Tornado and Sensitivity Charts Linear 496
136. Sensitivity—Tornado and Sensitivity Nonlinear 503

137. Simulation—Basic Simulation Model 510
138. Simulation—Best Surgical Team 517
139. Simulation—Correlated Simulation 525
140. Simulation—Correlation Effects on Risk 528
141. Simulation—Data Fitting 531
142. Simulation—Debt Repayment and Amortization 534
143. Simulation—Demand Curve and Elasticity Estimation 538
144. Simulation—Discounted Cash Flow, Return on Investment, and Volatility Estimates 542
145. Simulation—Infectious Diseases 546
146. Simulation—Recruitment Budget (Negative Binomial and Multidimensional Simulation) 548
147. Simulation—Retirement Funding with VBA Macros 556
148. Simulation—Roulette Wheel 560
149. Simulation—Time Value of Money 562

150. Six Sigma—Obtaining Statistical Probabilities, Basic Hypothesis Tests, Confidence Intervals, and Bootstrapping Statistics 571
151. Six Sigma—One- and Two-Sample Hypothesis Tests Using t-Tests, Z-Tests, F-Tests, ANOVA, and Nonparametric Tests (Friedman, Kruskal Wallis, Lilliefors, and Runs Tests) 590

152. Six Sigma—Sample Size Determination and Design of Experiments 623
153. Six Sigma—Statistical and Unit Capability Measures, Specification Levels, and Control Charts 627

154. Valuation—Buy versus Lease 631
155. Valuation—Banking: Classified Loan Borrowing Base 634
156. Valuation—Banking: Break-Even Inventory with Seasonal Lending Trial Balance Analysis 637
157. Valuation—Banking: Firm in Financial Distress 640
158. Valuation—Banking: Pricing Loan Fees Model 642
159. Valuation—Valuation Model 644

160. Value at Risk—Optimized and Simulated Portfolio VaR 647
161. Value at Risk—Options Delta Portfolio VaR 651
162. Value at Risk—Portfolio Operational and Credit Risk VaR Capital Adequacy 653
163. Value at Risk—Right-Tail Capital Requirements 657
164. Value at Risk—Static Covariance Method 661

165. Volatility—Implied Volatility 663
166. Volatility—Volatility Computations (Log Returns, Log Assets, Implied Volatility, Management Assumptions, EWMA, GARCH) 664

167. Yield Curve—CIR Model 673
168. Yield Curve—Curve Interpolation BIM Model 674
169. Yield Curve—Curve Interpolation NS Model 676
170. Yield Curve—Forward Rates from Spot Rates 678
171. Yield Curve—Term Structure of Volatility 679
172. Yield Curve—U.S. Treasury Risk-Free Rates and Cubic Spline Curves 680
173. Yield Curve—Vasicek Model 690

PART 2

Real Options SLS Applications 693

174. Introduction to the SLS Software 695
Single Asset and Single Phased Module 697
Multiple Asset or Multiple Phased SLS Module 704
Multinomial SLS Module 705
SLS Excel Solution Module 709
SLS Excel Functions Module 712
Lattice Maker Module 714

175. Employee Stock Options—Simple American Call Option 715
176. Employee Stock Options—Simple Bermudan Call Option with Vesting 716
177. Employee Stock Options—Simple European Call Option 719

178. Employee Stock Options—Suboptimal Exercise 720
179. Employee Stock Options—Vesting, Blackout, Suboptimal, Forfeiture 723

180. Exotic Options—American and European Lower Barrier Options 725
181. Exotic Options—American and European Upper Barrier Options 728
182. Exotic Options—American and European Double Barrier Options and Exotic Barriers 731
183. Exotic Options—Basic American, European, and Bermudan Call Options 734
184. Exotic Options—Basic American, European, and Bermudan Put Options 736

185. Real Options—American, European, Bermudan, and Customized Abandonment Options 739
186. Real Options—American, European, Bermudan, and Customized Contraction Options 749
187. Real Options—American, European, Bermudan, and Customized Expansion Options 756
188. Real Options—Contraction, Expansion, and Abandonment Options 763
189. Real Options—Dual Variable Rainbow Option Using Pentanomial Lattices 767
190. Real Options—Exotic Chooser Options 770
191. Real Options—Exotic Complex Floating American and European Chooser 771
192. Real Options—Jump-Diffusion Option Using Quadranomial Lattices 774
193. Real Options—Mean-Reverting Calls and Puts Using Trinomial Lattices 777
194. Real Options—Multiple Assets Competing Options 779
195. Real Options—Path-Dependent, Path-Independent, Mutually Exclusive, Non–Mutually Exclusive, and Complex Combinatorial Nested Options 781
196. Real Options—Sequential Compound Options 783
197. Real Options—Simultaneous Compound Options 791
198. Real Options—Simple Calls and Puts Using Trinomial Lattices 795

PART 3

Real Options Strategic Case Studies—Framing the Options 799

199. Real Options Strategic Cases—High-Tech Manufacturing: Build or Buy Decision with Real Options 801
200. Real Options Strategic Cases—Oil and Gas: Farm-Outs, Options to Defer, and Value of Information 810

201. Real Options Strategic Cases—Pharmaceutical Development: Value of Perfect Information and Optimal Trigger Values 814
202. Real Options Strategic Cases—Option to Switch Inputs 817
203. Valuation—Convertible Warrants with a Vesting Period and Put Protection 821

APPENDIX A
List of Models **827**

APPENDIX B
List of Functions **837**

APPENDIX C
Understanding and Choosing the Right Probability Distributions **899**

APPENDIX D
Financial Statement Analysis **919**

APPENDIX E
Exotic Options Formulae **927**

APPENDIX F
Measures of Risk **941**

APPENDIX G
Mathematical Structures of Stochastic Processes **957**

Glossary of Input Variables and Parameters in the Modeling Toolkit Software **963**

About the DVD **995**

About the Author **999**

Index **1001**

Preface

Advanced Analytical Models is a large collection of advanced models with a multitude of industry and domain applications. The book is based on years of academic research and practical consulting experience, coupled with domain expert contributions. The Modeling Toolkit software that holds all the models, Risk Simulator software, and Real Options SLS software were all developed by the author, with over 1,000 functions, tools, and model templates in these software applications. The trial versions are included in the accompanying DVD.

The applications covered are vast. Included are Basel II banking risk requirements (credit risk, market risk, credit spreads, default risk, value at risk, etc.) and financial analysis (exotic options and valuation), risk analysis (stochastic forecasting, risk-based Monte Carlo simulation, optimization), real options analysis (strategic options and decision analysis), Six Sigma and quality initiatives, management science and statistical applications, and everything in between, such as applied statistics, manufacturing, decision analysis, operations research, optimization, forecasting, and econometrics.

This book is targeted at practitioners who require the algorithms, examples, models, and insights in solving more advanced and even esoteric problems. This book does not only talk about modeling or illustrate basic concepts and examples; it comes complete with a DVD filled with sample modeling videos, case studies, and software applications to help you get started immediately. In other words, this book dispenses with all the theoretical discussions and mathematical models that are extremely hard to decipher and apply in the real business world. Instead, these theoretical models have been coded up into user-friendly and powerful software, and this book shows the reader how to start applying advanced modeling techniques almost immediately. The trial software applications allow you to access the approximately 300 model templates and 800 functions and tools, understand the concepts, and use embedded functions and algorithms in their own models. In addition, you can run risk-based Monte Carlo simulations and advanced forecasting methods, and perform optimization on a myriad of situations, as well as structure and solve customized real options and financial options problems.

Each model template that comes in the Modeling Toolkit software is described in this book. Descriptions are provided in as much detail as the applications warrant. Some of the more fundamental concepts in risk analysis and real options are covered in the author's other books. It is suggested that these books, *Modeling Toolkit: Applying Monte Carlo Simulation, Real Options Analysis, Stochastic Forecasting, and Portfolio Optimization* (2006) and *Real Options Analysis, Second Edition* (2005),

both published by John Wiley & Sons, be used as references for some of the models in this book. Those modeling issues that are, in the author's opinion, critical, whether they are basic issues or more advanced analytical ones, are presented in detail. As software applications change continually, it is recommended that you check the author's web site (www.realoptionsvaluation.com) frequently for any analytical updates, software upgrades, and revised or new models.

ACKNOWLEDGMENTS

A special thank you to the contributors, including Mark Benyovszky, Morton Glantz, Uriel Kusiatin, and Victor Wong.

Dr. Johnathan Mun
JohnathanMun@cs.com
California, USA

Software Applications

This book covers the following software applications:

Modeling Toolkit

Over 800 functions, models, and tools and over 300 Excel and SLS templates covering the following applications:

- Business analytics and statistics (CDF, ICDF, PDF, data analysis, integration)
- Credit and Debt Analysis (credit default swap, credit spread options, credit rating, debt options and pricing)
- Decision Analysis (decision tree, Minimax, utility functions)
- Exotic Options (over 100 types of financial and exotic options)
- Forecasting (ARIMA, econometrics, EWMA, GARCH, nonlinear extrapolation, spline, time-series)
- Industry Applications (banking, biotech, insurance, IT, real estate, utility)
- Operations Research and Portfolio Optimization (continuous, discrete, integer, static, dynamic, and stochastic)
- Options Analysis (BDT interest lattices, debt options, options trading strategies)
- Portfolio Models (investment allocations, optimization, risk and return profiles)
- Probability of Default and Banking Credit Risk (private, public and retail debt, credit derivatives and swaps)
- Real Options Analysis (over 100 types: abandon, barrier, contract, customized, dual asset, expand, multi-asset, multi-phased, pentanomials, quadranomials, sequential, switch, and trinomials)
- Risk Hedging (delta and delta-gamma hedges, foreign exchange and interest rate risk)
- Risk Simulation (correlated simulation, data fitting, Monte Carlo simulation, risk-simulation)
- Six Sigma (capability measures, control charts, hypothesis tests, measurement systems, precision, sample size)
- Statistical Tools (ANOVA, Two-Way ANOVA, nonparametric hypotheses tests, parametric tests, principal components, variance-covariance)
- Valuation (APT, buy versus lease, CAPM, caps and floors, convertibles, financial ratios, valuation models)
- Value at Risk (static covariance and simulation-based VaR)
- Volatility (EWMA, GARCH, implied volatility, Log Returns, Real Options Volatility, probability to volatility)
- Yield Curve (BIS, Cox, Merton, NS, spline, Vasicek)

Risk Simulator

Over 25 statistical distributions covering the following applications:

- Applied Business Statistics (descriptive statistics, CDF/ICDF/PDF probabilities, stochastic parameter calibration)
- Bootstrap Nonparametric Simulation and Hypothesis Testing (testing empirical and theoretical moments)
- Correlated Simulations (simulation copulas and Monte Carlo)
- Data Analysis and Regression Diagnostics (heteroskedasticity, multicollinearity, nonlinearity, outliers)
- Forecasting (ARIMA, Auto-ARIMA, J-S curves, GARCH, Markov chains, multivariate regressions, stochastic processes)
- Optimization (static, dynamic, stochastic)
- Sensitivity Analysis (correlated sensitivity, scenario, spider, tornado)

Real Options SLS

Customizable Binomial, Trinomial, Quadranomial, and Pentanomial Lattices

Lattice Makers (lattices with Monte Carlo simulation)

Super fast super lattice algorithms (running thousands of lattice steps in seconds)

Covering the following applications:

- Exotic Options Models (barriers, benchmarked, multiple assets, portfolio options)
- Financial Options Models (3D dual asset exchange, single and double barriers)
- Real Options Models (abandon, barrier, contract, expand, sequential compound, switching)
- Specialized Options (mean-reverting, jump-diffusion, and dual asset rainbows)

Employee Stock Options Valuation Toolkit

Applied by the U.S. Financial Accounting Standards Board for FAS 123R 2004

Binomial and closed-form models

Covers:

- Blackout Periods
- Changing Volatility
- Forfeiture Rates
- Suboptimal Exercise Multiple
- Vesting

PART 1

Modeling Toolkit and Risk Simulator Applications

This book covers about 300 different analytical model templates that apply up to 800 modeling functions and tools from a variety of software applications. Trial versions of these software applications are included in the book's DVD or can be downloaded directly from the Web at www.realoptionsvaluation.com. Part I of the book deals with models using the Modeling Toolkit and Risk Simulator software applications. Part II deals with real options and financial option models using the Real Options SLS software. Readers who are currently expert users of the Modeling Toolkit software and Risk Simulator software may skip this section and dive directly into the models.

INTRODUCTION TO THE MODELING TOOLKIT SOFTWARE

The Modeling Toolkit software incorporates about 800 different advanced analytical models, functions, and tools, applicable in a variety of industries and applications. Appendix 1 lists the models available in the software as of this book's publication date. To install this software for a trial period of 30 days, insert the DVD that comes with the book or visit www.realoptionsvaluation.com and click on **Downloads**. Look for the Modeling Toolkit software. This software works on Windows XP, or Vista, and requires Excel XP, 2003 or 2007 to run. At the end of the installation process, you will be prompted for a license key. Please use this trial license:

Name: **30 Day Trial**
Key: 4C55-0BA2-420E-CA84

To start the software, click on **Start | Programs | Real Options Valuation | Modeling Toolkit | Modeling Toolkit**. This action will start Excel. Inside Excel, you will notice a new menu item called **Modeling Toolkit**. This menu is self-explanatory, as the models are categorized by application domain, and each model is described in more detail in this book. Please note that this software uses Excel macros. If you receive an error message on macros, it is because your system is set to a high security level. You need to fix this by starting Excel XP or 2003 and clicking on **Tools | Macros | Security | Medium** and restarting the software. If you are using Excel 2007, you can simply click on **Enable Macros** when prompted (or reset your security settings when in Excel 2007 by clicking on the *Office* button located at

the top left of the screen and selecting **Excel Options | Trust Center | Trust Center Settings | Macro Settings | Enable All Macros**).

Note that the trial version will expire in 30 days. To obtain a full corporate license, please contact the author's firm, Real Options Valuation, Inc., at admin@realoptionsvaluation.com or visit the company's web site (www.realoptionsvaluation.com). Notice that after the software expiration date, some of the models that depend on Risk Simulator or Real Options SLS software still will function, until their respective expiration dates. In addition, after the expiration date, these worksheets still will be visible, but the analytical results and functions will return null values. Finally, software versions continually change and improve, and the best recommendation is to visit the company's web site for any updated or newer software versions or details on installation and licensing.

The Appendixes provide a more detailed list of all the functions, tools, and models and the Glossary details the required variable inputs in this software.

INTRODUCTION TO RISK SIMULATOR

This section also provides the novice risk analyst an introduction to the Risk Simulator software for performing Monte Carlo simulation, where a trial version of the software is included in the book's DVD. Please refer to About the DVD at the end of this book for details on obtaining this extended trial license. This section starts off by illustrating what Risk Simulator does and what steps are taken in a Monte Carlo simulation as well as some of the more basic elements in a simulation analysis. It continues with how to interpret the results from a simulation and ends with a discussion of correlating variables in a simulation as well as applying precision and error control. Many more advanced techniques such as ARIMA forecasts and optimization are also discussed. Software versions with new enhancements are released continually. Please review the software's user manual and the software download site (www.realoptionsvaluation.com) for more up-to-date details on using the latest version of the software. See *Modeling Risk: Applying Monte Carlo Simulation, Real Options Analysis, Stochastic Forecasting, and Portfolio Optimization* (Hoboken, NJ: John Wiley & Sons, 2007), also by the author, for more technical details on using Risk Simulator.

Risk Simulator is a Monte Carlo simulation, forecasting, and optimization software. It is written in Microsoft .NET C# and functions with Excel as an add-in. This software is compatible and often used with the Real Options SLS software shown in Part II of this book, also developed by the author. Stand-alone software applications in C++ are also available for implementation into other existing proprietary software or databases.

The different functions or modules in both software applications are briefly described next. The Appendixes provide a more detailed list of all the functions, tools, and models.

- The *Simulation Module* allows you to:
  - Run simulations in your existing Excel-based models
  - Generate and extract simulation forecasts (distributions of results)
  - Perform distributional fitting (automatically finding the best-fitting statistical distribution)

- Compute correlations (maintaining relationships among multiple simulated random variables)
- Identify sensitivities (creating tornado and sensitivity charts)
- Test statistical hypotheses (finding statistical differences and similarities between pairs of forecasts)
- Run bootstrap simulation (testing the robustness of result statistics)
- Run custom and nonparametric simulations (simulations using historical data without specifying any distributions or their parameters for forecasting without data or applying expert opinion forecasts)

- The *Forecasting Module* can be used to generate:
  - Automatic time-series forecasts (with and without seasonality and trend)
  - Automatic ARIMA (automatically generate the best-fitting ARIMA forecasts)
  - Basic Econometrics (modified multivariate regression forecasts)
  - Box-Jenkins ARIMA (econometric forecasts)
  - GARCH Models (forecasting and modeling volatility)
  - J-Curves (exponential growth forecasts)
  - Markov Chains (market share and dynamics forecasts)
  - Multivariate regressions (modeling linear and nonlinear relationships among variables)
  - Nonlinear extrapolations (curve fitting)
  - S-Curves (logistic growth forecasts)
  - Spline Curves (interpolating and extrapolating missing nonlinear values)
  - Stochastic processes forecasts (random walks, mean-reversions, jump-diffusions, and mixed processes).
- The *Optimization* module is used for running:
  - Linear and nonlinear optimization
  - Static optimization (without simulation), dynamic optimization (with simulation), and stochastic optimization (with simulation and run multiple times)
  - Discrete, continuous and integer decision variables
- *Analytical Tools*
  - Correlated simulations
  - Data diagnostics (autocorrelation, correlation, distributive lags, heteroskedasticity, micronumerosity, multicollinearity, nonlinearity, nonstationarity, normality, outliers, partial autocorrelation, and others)
  - Data extraction
  - Data fitting
  - Data import and export
  - Distribution analysis (PDF, CDF, ICDF)
  - Distribution designer (creating customized distributions and Delphi simulation)
  - Hypothesis tests and bootstrap simulation
  - Sensitivity and dynamic scenario analysis
  - Statistical analysis (descriptive statistics, distributional fitting, hypothesis tests, nonlinear extrapolation, normality, stochastic parameter estimation, time-series forecasts, trending, and others)
  - Tornado and spider charts
- The Real Options Super Lattice Solver (SLS) is another stand-alone software that complements Risk Simulator, used for solving simple to complex real options problems. See Part II of this book for details on this software's applications.

To install the software, insert the accompanying DVD, click on the *Install Risk Simulator* link, and follow the onscreen instructions. You will need to be online to download the latest version of the software. The software requires Windows XP/Vista, administrative privileges, and Microsoft .NET Framework 1.1 and 2.0 installed on the computer. Most new computers come with Microsoft .NET Framework 1.1 already preinstalled. However, if an error message pertaining to requiring .NET Framework occurs during the installation of Risk Simulator, exit the installation. Then install the relevant .NET Framework software, also included in the DVD (found in the *DOT NET Framework* folder). Complete the .NET installation, restart the computer, and then reinstall the Risk Simulator software. Version 1.1 of the .NET Framework is required even if your system has version 2.0/3.0, as they work independently of each other. You may also download this software on the Download page of www.realoptionsvaluation.com. See the About the DVD section at the end of this book for details on obtaining an extended trial license.

Once installation is complete, start Microsoft Excel. If the installation was successful, you should see an additional *Risk Simulator* item on the menu bar in Excel and a new icon bar, as shown in Figure I.1. Figure I.2 shows the icon toolbar in more detail.

Please note that Risk Simulator supports multiple languages (e.g., English, Chinese, Japanese, and Spanish) and you can switch among languages by going to **Risk Simulator | Languages**.

You are now ready to start using the software for a trial period. You can obtain permanent or academic licenses from www.realoptionsvaluation.com.

If you are using Windows Vista, make sure to disable User Access Control before installing the software license. To do so: Click on **Start | Control Panel | Classic View**

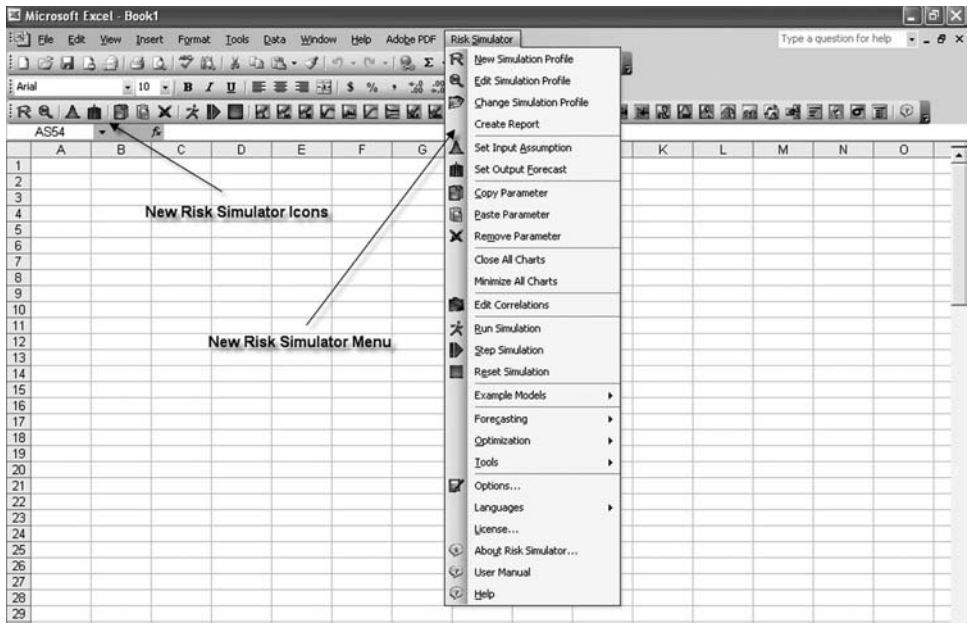


FIGURE I.1 Risk Simulator menu and icon toolbar (*Continued*)

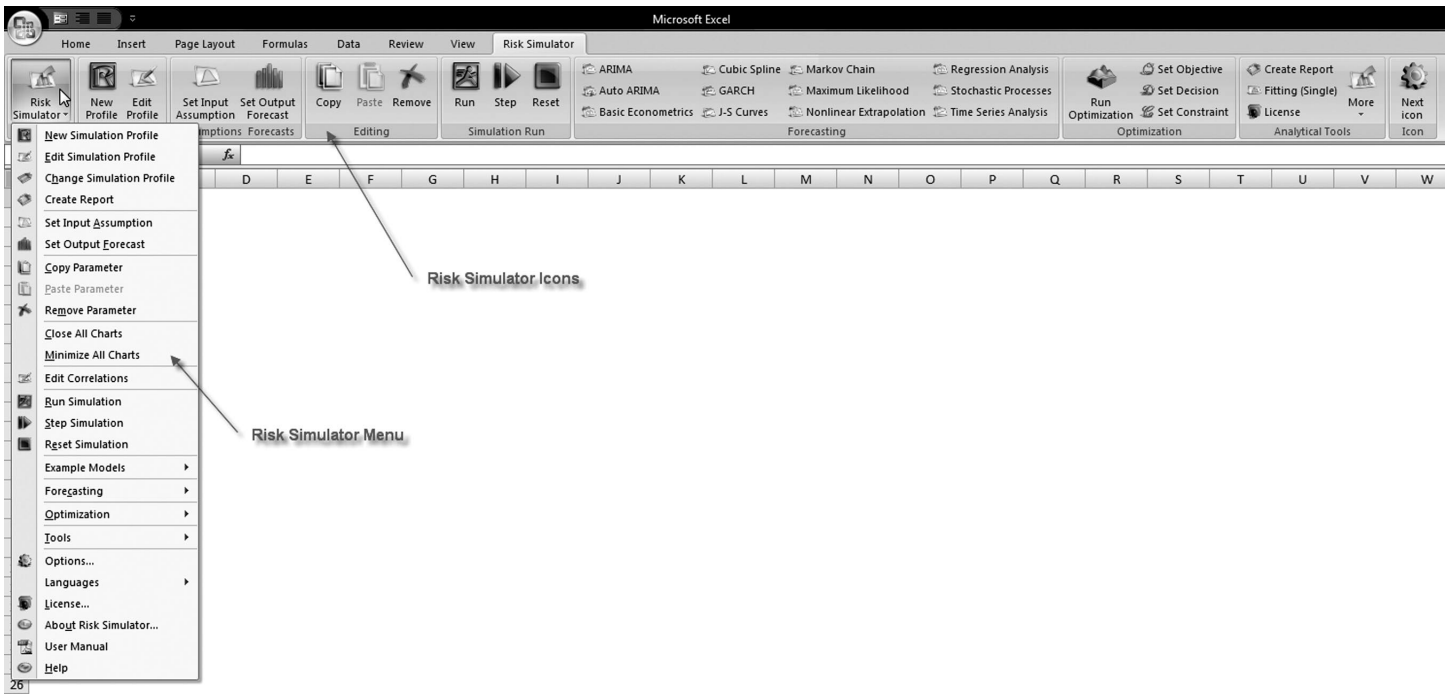


FIGURE I.1 (*Continued*)

(**on the left panel**) | **User Accounts** | **Turn User Account Control On or Off** and uncheck the option, **Use User Account Control** (**UAC**), and restart the computer. When restarting the computer, you will get a message that UAC is turned off. You can turn this message off by going to the **Control Panel** | **Security Center** | **Change the Way Security Center Alerts Me** | **Don't Notify Me and Don't Display the Icon.**

The sections that follow provide step-by-step instructions for using the software. As the software is continually updated and improved, the examples in this book might be slightly different from the latest version downloaded from the Internet.

RUNNING A MONTE CARLO SIMULATION

Typically, to run a simulation in your existing Excel model, you must perform these five steps:

1. Start a new or open an existing simulation profile.
2. Define input assumptions in the relevant cells.
3. Define output forecasts in the relevant cells.
4. Run the simulation.
5. Interpret the results.

If desired, and for practice, open the example file called *Basic Simulation Model* and follow along the examples on creating a simulation. The example file can be found on the menu at **Risk Simulator** | **Example Models.**

1. Starting Models a New Simulation Profile

To start a new simulation, you must first create a simulation profile. A simulation profile contains a complete set of instructions on how you would like to run a simulation; it contains all the assumptions, forecasts, simulation run preferences, and so forth. Having profiles facilitates creating multiple scenarios of simulations; that is, using the same exact model, several profiles can be created, each with its own specific simulation assumptions, forecasts, properties, and requirements. The same analyst can create different test scenarios using different distributional assumptions and inputs or multiple users can test their own assumptions and inputs on the same model. Instead of having to make duplicates of the model, the same model can be used and different simulations can be run through this model *profiling* process.

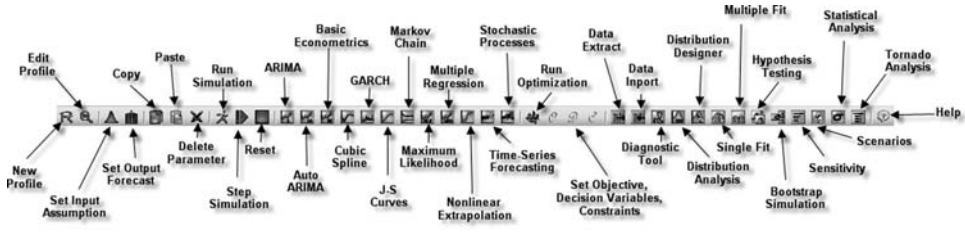


FIGURE 1.2 Risk Simulator icon toolbar (*Continued*)

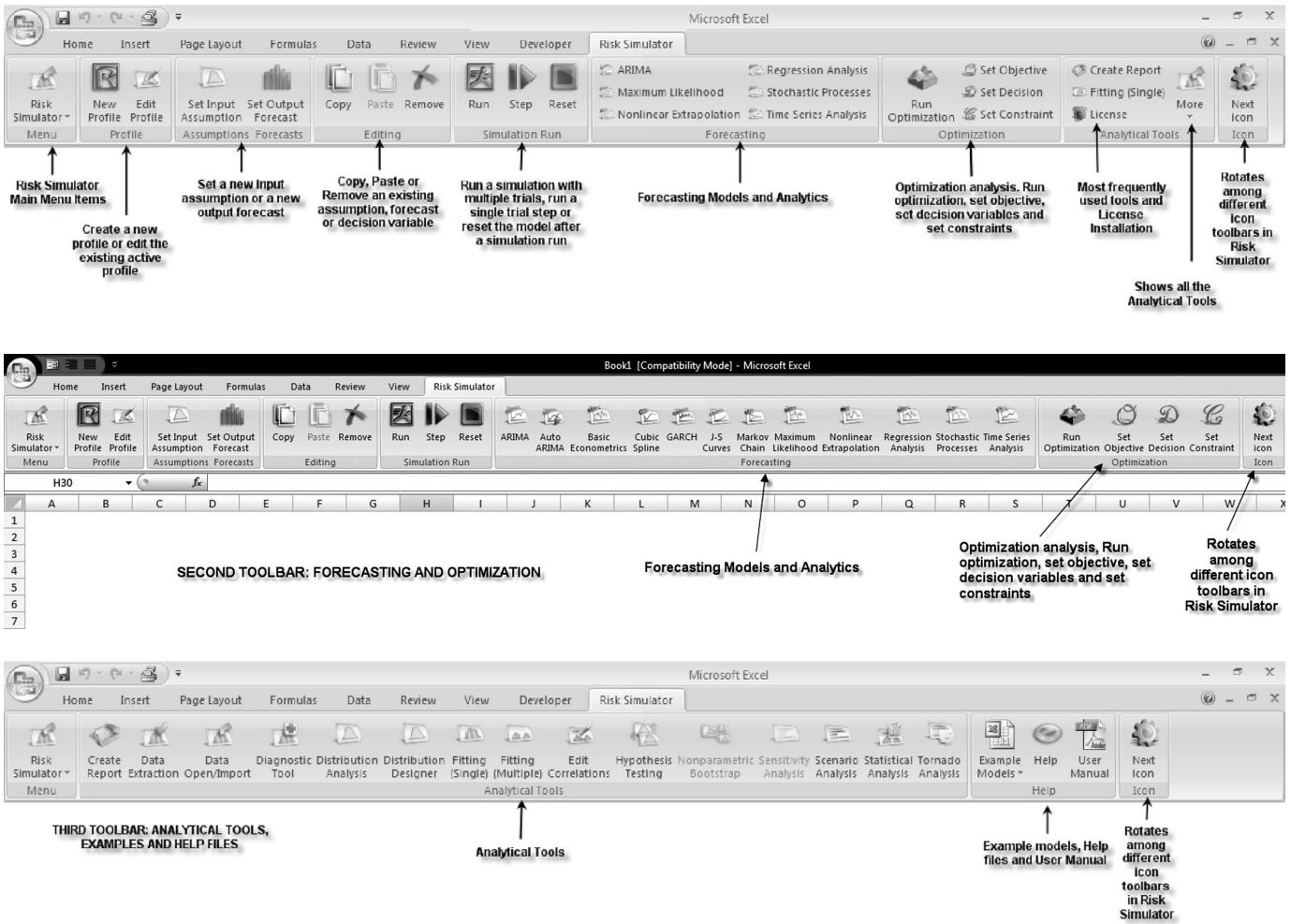


FIGURE I.2 (*Continued*)

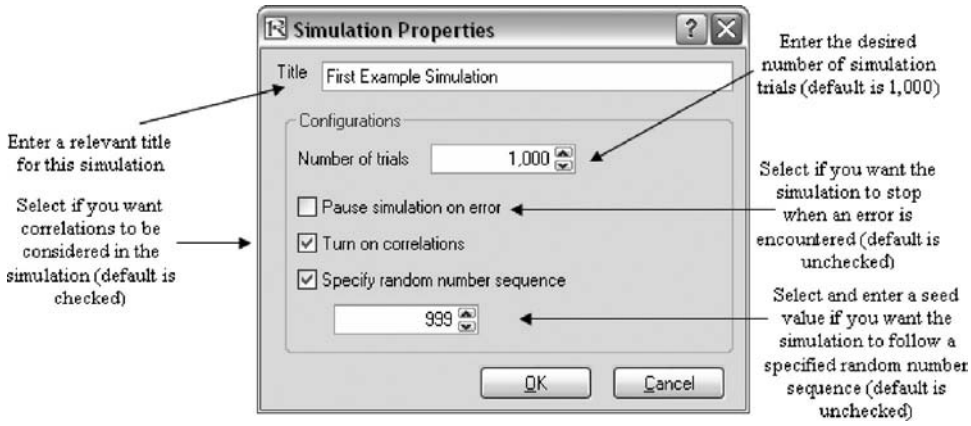

Simulation Properties
Title
First Example Simulation
Configurations
Number of trials
1,000
Pause simulation on error
Turn on correlations
Specify random number sequence
999
OK
Cancel
Enter a relevant title for this simulation
Select if you want correlations to be considered in the simulation (default is checked)
Enter the desired number of simulation trials (default is 1,000)
Select if you want the simulation to stop when an error is encountered (default is unchecked)
Select and enter a seed value if you want the simulation to follow a specified random number sequence (default is unchecked)

FIGURE I.3 New simulation profile

Start a new simulation profile by performing these three steps.

1. **Start Excel** and create a new or open an existing model. You can use the Basic Simulation Model example to follow along: **Risk Simulator | Examples | Basic Simulation Model.**
2. Click on **Risk Simulator | New Simulation Profile**.
3. Enter a title for your simulation including all other pertinent information (Figure I.3).

The elements in the new simulation profile dialog, shown in Figure I.3, include:

- *Title*. Specifying a simulation profile name or title allows you to create multiple simulation profiles in a single Excel model. By so doing, you can save different simulation scenario profiles within the same model without having to delete existing assumptions and change them each time a new simulation scenario is required.
- *Number of trials*. Enter the number of simulation trials required. Running 1,000 trials means that 1,000 different iterations of outcomes based on the input assumptions will be generated. You can change this number as desired, but the input has to be positive integers. The default number of runs is 1,000 trials.
- *Pause on simulation error*. If checked, the simulation stops every time an error is encountered in the Excel model; that is, if your model encounters a computational error (e.g., some input values generated in a simulation trial may yield a divide-by-zero error in a spreadsheet cell), the simulation stops. This feature is important to help audit your model to make sure there are no computational errors in your Excel model. However, if you are sure the model works, there is no need for you to check this preference.
- *Turn on correlations*. If checked, correlations between paired input assumptions will be computed. Otherwise, correlations will all be set to zero and a simulation is run assuming no cross-correlations between input assumptions. Applying correlations will yield more accurate results if correlations do indeed exist and will tend to yield a lower forecast confidence if negative correlations exist.

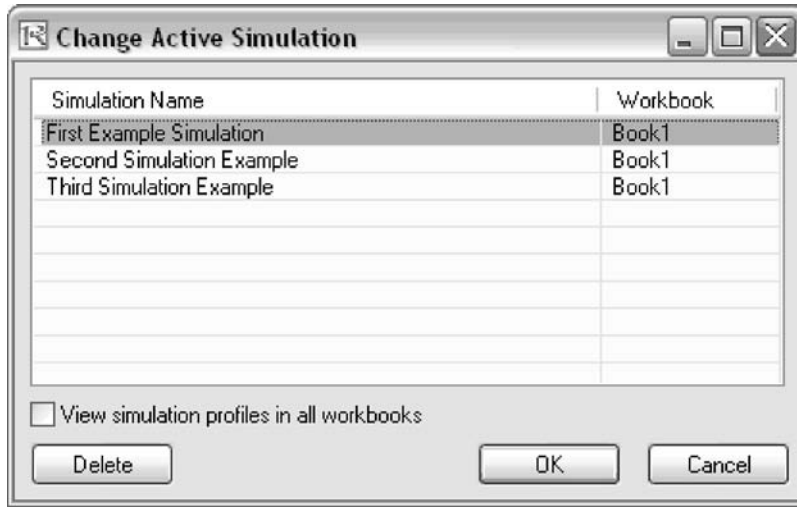


FIGURE I.4 Change active simulation

- *Specify random number sequence.* By definition, a simulation yields slightly different results every time it is run by virtue of the random-number generation routine in Monte Carlo simulation. This is a theoretical fact in all random-number generators. However, when making presentations, sometimes you may require the same results. (For example, during a live presentation you may like to show the same results that are in some pregenerated printed reports from a previous simulation run; when you are sharing models with others, you also may want the same results to be obtained every time.) If that is the case, check this preference and enter in an initial seed number. The seed number can be any positive integer. Using the same initial seed value, the same number of trials, and the same input assumptions always will yield the same sequence of random numbers, guaranteeing the same final set of results.

Note that once a new simulation profile has been created, you can come back later and modify your selections. In order to do this, make sure that the current active profile is the profile you wish to modify; otherwise, click on **Risk Simulator | Change Simulation Profile,** select the profile you wish to change, and click **OK.** (Figure I.4 shows an example where there are multiple profiles and how to activate, duplicate or delete a selected profile.) Then click on **Risk Simulator | Edit Simulation Profile** and make the required changes.

2. Defining Input Assumptions

The next step is to set input assumptions in your model. Note that assumptions can be assigned only to cells without any equations or functions (i.e., typed-in numerical values that are inputs in a model), whereas output forecasts can be assigned only to cells with equations and functions (i.e., outputs of a model). Recall that assumptions

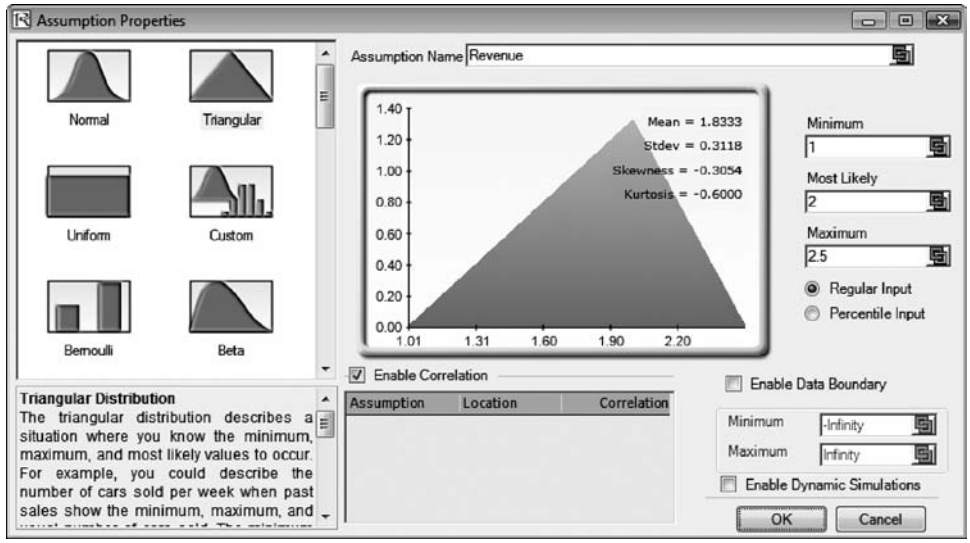


FIGURE I.5 Setting an input assumption

and forecasts cannot be set unless a simulation profile already exists. Follow these three steps to set new input assumptions in your model:

1. Select the cell you wish to set an assumption on (e.g., cell G8 in the Basic Simulation Model example).
2. Click on **Risk Simulator | Set Input Assumption** or click the *Set Assumption* icon in the Risk Simulator icon toolbar.
3. Select the relevant distribution you want, enter the relevant distribution parameters, and hit **OK** to insert the input assumption into your model (Figure I.5).

Several key areas are worthy of mention in the Assumption Properties. Figure I.6 shows the different areas.

- *Assumption Name*. This optional area allows you to enter in unique names for the assumptions to help track what each of the assumptions represents. Good modeling practice is to use short but precise assumption names.
- *Distribution Gallery*. This area shows all of the different distributions available in the software. To change the views, right-click anywhere in the gallery and select large icons, small icons, or list. More than two dozen distributions are available.
- *Input Parameters*. Depending on the distribution selected, the required relevant parameters are shown. You may either enter the parameters directly or link them to specific cells in your worksheet. Click on the *Link* icon to link an input parameter to a worksheet cell. Hard-coding or typing the parameters is useful when the assumption parameters are assumed not to change. Linking to worksheet cells is useful when the input parameters themselves need to be visible on the worksheets or can be changed, as in a dynamic simulation (where the input parameters themselves are linked to assumptions in the worksheets creating a multidimensional simulation, or simulation of simulations).

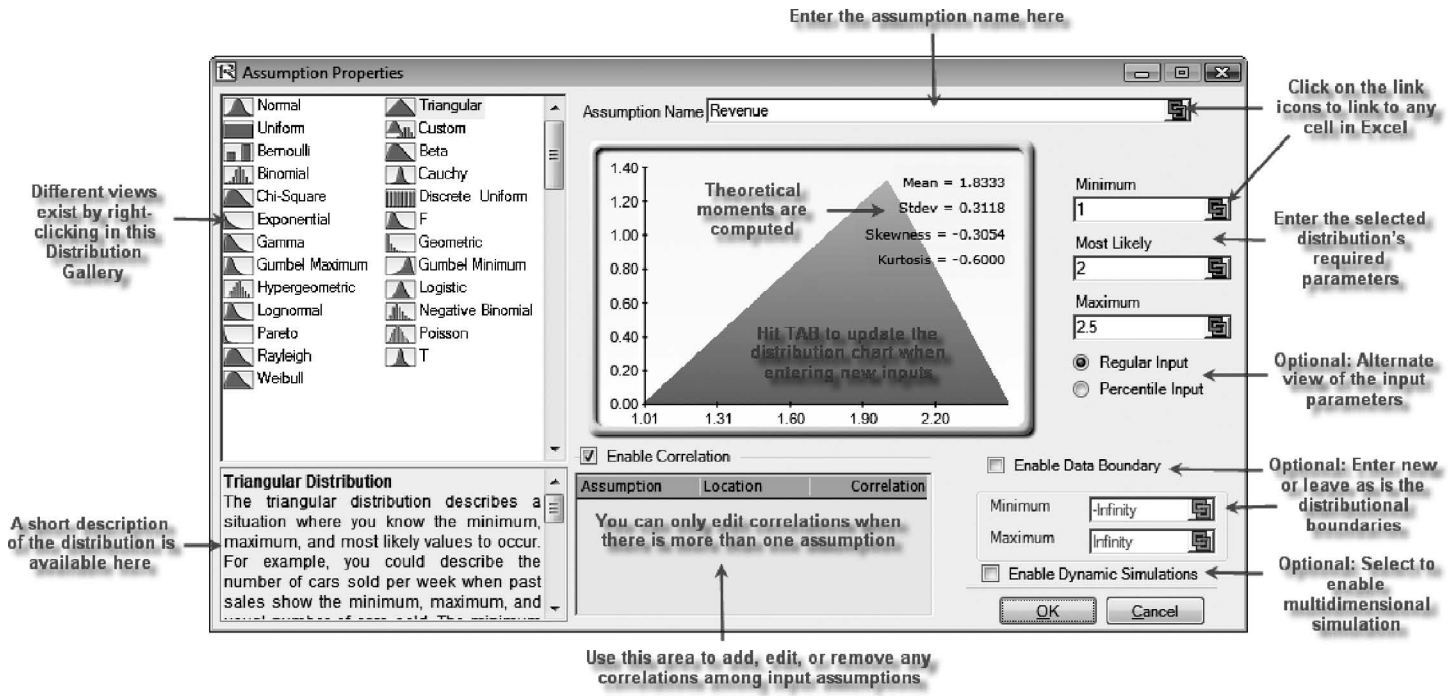


FIGURE 1.6 Assumption properties

- *Data Boundary.* Typically, the average analyst does not use distributional or data boundaries truncation, but they exist for truncating the distributional assumptions. For instance, if a normal distribution is selected, the theoretical boundaries are between negative infinity and positive infinity. However, in practice, the simulated variable exists only within some smaller range. This range can be entered to truncate the distribution appropriately.
- *Correlations.* Pairwise correlations can be assigned to input assumptions here. If assumptions are required, remember to check the *Turn on Correlations* preference by clicking on **Risk Simulator | Edit Simulation Profile.** See the discussion on correlations later in this chapter for more details about assigning correlations and the effects correlations will have on a model.
- *Short Descriptions.* Short descriptions exist for each of the distributions in the gallery. The short descriptions explain when a certain distribution is used as well as the input parameter requirements. See the section in the appendix "Understanding Probability Distributions" in *Modeling Risk: Applying Monte Carlo Simulation, Real Options Analysis, Stochastic Forecasting, and Portfolio Optimization* (Hoboken, NJ: John Wiley & Sons, 2006), also by the author, for details about each distribution type available in the software.

Note: If you are following along with the example, continue by setting another assumption on cell G9. This time use the Uniform distribution with a minimum value of 0.9 and a maximum value of 1.1. Then proceed to defining the output forecasts in the next step.

3. Defining Output Forecasts

The next step is to define output forecasts in the model. Forecasts can be defined only on output cells with equations or functions.

Use these three steps to define the forecasts:

1. Select the cell on which you wish to set an assumption (e.g., cell G10 in the Basic Simulation Model example).
2. Click on **Risk Simulator | Set Output Forecast** or click on the set forecast icon on the Risk Simulator icon toolbar.
3. Enter the relevant information and click **OK**.

Figure I.7 illustrates the set forecast properties, which include:

- *Forecast Name.* Specify the name of the forecast cell. This is important because when you have a large model with multiple forecast cells, naming the forecast cells individually allows you to access the right results quickly. Do not underestimate the importance of this simple step. Good modeling practice is to use short but precise assumption names.
- *Forecast Precision.* Instead of relying on a guesstimate of how many trials to run in your simulation, you can set up precision and error controls. When an error-precision combination has been achieved in the simulation, the simulation will pause and inform you of the precision achieved. Thus the number of simulation trials is an automated process; you do not have to guess the required number

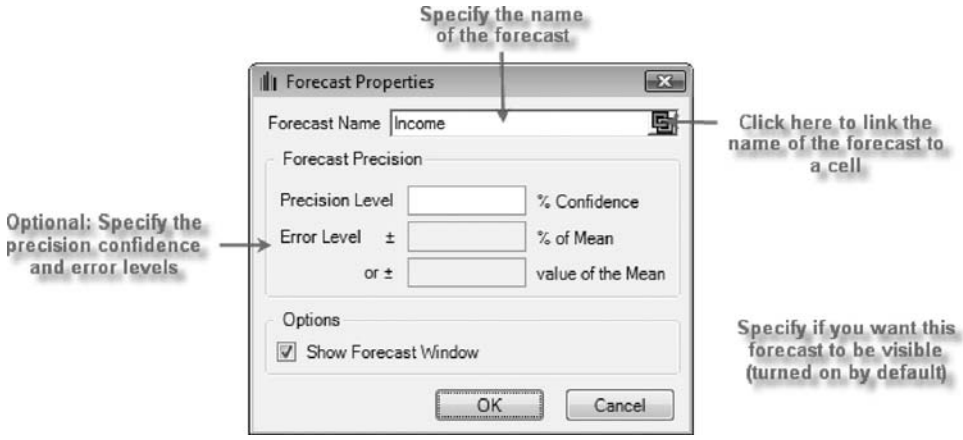


FIGURE I.7 Set output forecast

of trials to simulate. Review the section on error and precision control for more specific details.

- *Show Forecast Window*. This property allows you to show or not show a particular forecast window. The default is to always show a forecast chart.

4. Run Simulation

If everything looks right, simply click on **Risk Simulator | Run Simulation** or click on the *Run* icon on the Risk Simulator toolbar, and the simulation will proceed. You may also reset a simulation after it has run to rerun it (**Risk Simulator | Reset Simulation** or the *Reset* icon on the toolbar), or to pause it during a run. Also, the step function (**Risk Simulator | Step Simulation** or the *Step* icon on the toolbar) allows you to simulate a single trial, one at a time, which is useful for educating others on simulation (i.e., you can show that at each trial, all the values in the assumption cells are being replaced and the entire model is recalculated each time).

5. Interpreting the Forecast Results

The final step in Monte Carlo simulation is to interpret the resulting forecast charts. Figures I.8 to I.15 show the forecast chart and the statistics generated after running the simulation. Typically, these sections on the forecast window are important in interpreting the results of a simulation:

- *Forecast Chart*. The forecast chart shown in Figure I.8 is a probability histogram that shows the frequency counts of values occurring and the total number of trials simulated. The vertical bars show the frequency of a particular x value occurring out of the total number of trials, while the cumulative frequency (smooth line) shows the total probabilities of all values at and below x occurring in the forecast.
- *Forecast Statistics*. The forecast statistics shown in Figure I.9 summarize the distribution of the forecast values in terms of the four moments of a distribution.

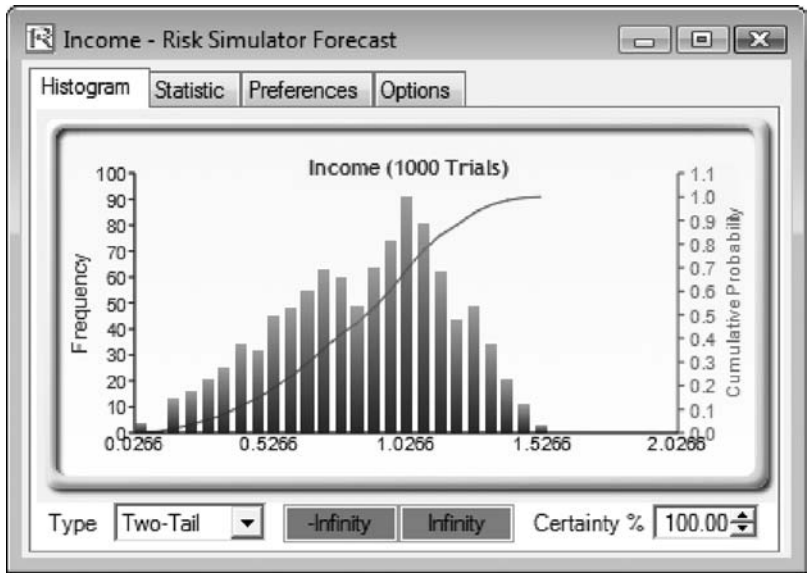
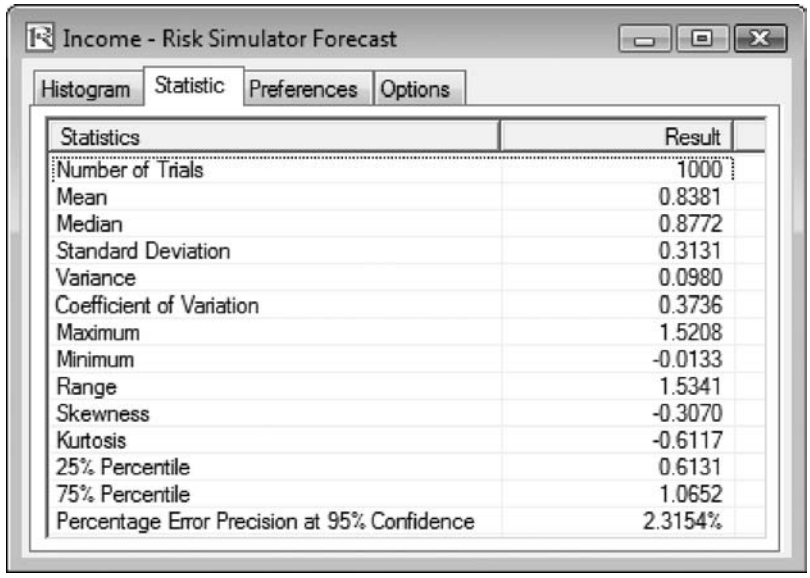


FIGURE I.8 Forecast chart

You can rotate between the histogram and statistics tab by depressing the space bar.

- *Preferences*. The preferences tab in the forecast chart (Figure I.10) allows you to change the look and feel of the charts. For instance, if *Always Show Window On Top* is selected, the forecast charts will always be visible regardless of what other software is running on your computer. The *Semitransparent When Inactive* is a powerful option used to compare or overlay multiple forecast charts at once



Statistics	Result
Number of Trials	1000
Mean	0.8381
Median	0.8772
Standard Deviation	0.3131
Variance	0.0980
Coefficient of Variation	0.3736
Maximum	1.5208
Minimum	-0.0133
Range	1.5341
Skewness	-0.3070
Kurtosis	-0.6117
25% Percentile	0.6131
75% Percentile	1.0652
Percentage Error Precision at 95% Confidence	2.3154%

FIGURE I.9 Forecast statistics

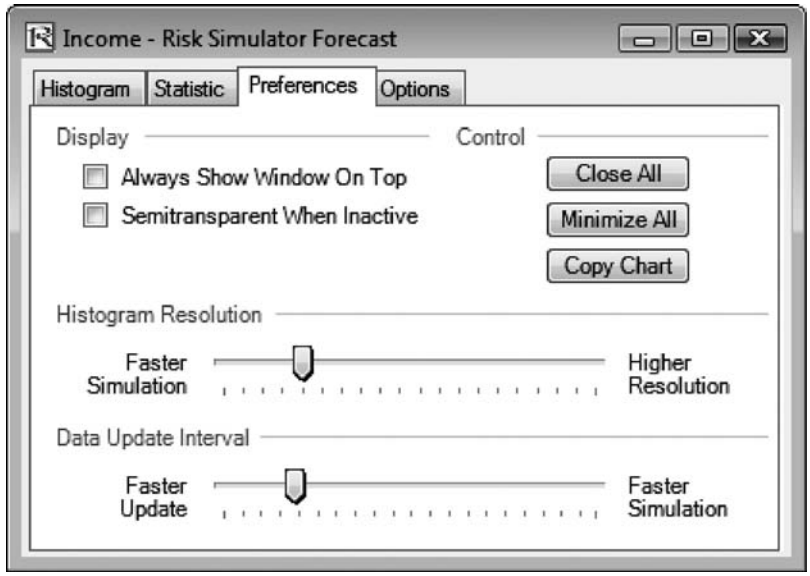


FIGURE I.10 Forecast chart preferences

(e.g., enable this option on several forecast charts and drag them on top of one another to visually see the similarities or differences). *Histogram Resolution* allows you to change the number of bins of the histogram, anywhere from 5 bins to 100 bins. Also, the *Update Data Interval* section allows you to control how fast the simulation runs versus how often the forecast chart is updated. That is, if you wish to see the forecast chart updated at almost every trial, this will slow down the simulation as more memory is being allocated to updating

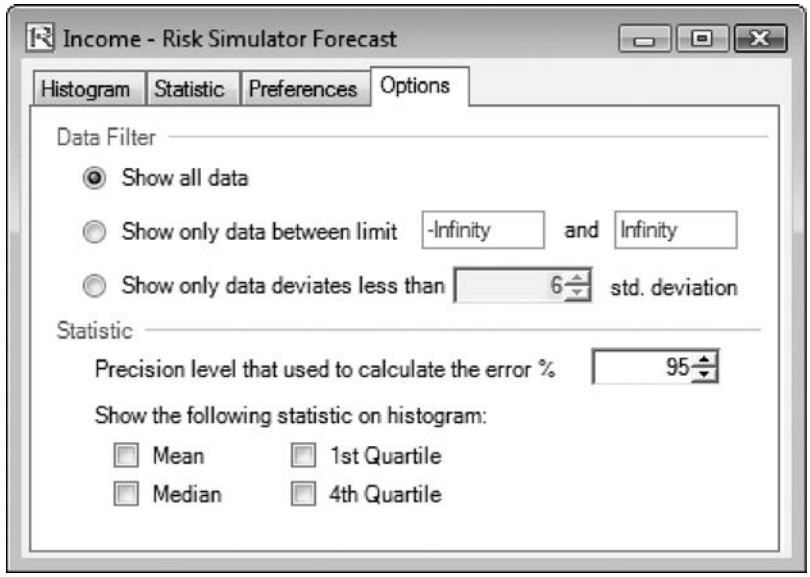


FIGURE I.11 Forecast chart options

the chart versus running the simulation. This is merely a user preference and in no way changes the results of the simulation, just the speed of completing the simulation. The *Copy Chart* button allows you to copy the active forecast chart for pasting into other software applications (e.g., PowerPoint or Word) and the *Close All* and *Minimize All* buttons allow you to control all opened forecast charts at once.

- *Options*. This forecast chart option (Figure I.11) allows you to show all the forecast data or to filter in or out values that fall within some specified interval or within some standard deviation that you choose. Also, you can set the precision level here for this specific forecast to show the error levels in the statistics view. See the section on precision and error control for more details.

USING FORECAST CHARTS AND CONFIDENCE INTERVALS

In forecast charts, you can determine the probability of occurrence called *confidence intervals*; that is, given two values, what are the chances that the outcome will fall between these two values? Figure I.12 illustrates that there is a 90% probability that the final outcome (in this case, the level of income) will be between \$0.2781 and \$1.3068. The two-tailed confidence interval can be obtained by first selecting **Two-Tail** as the type, entering the desired certainty value (e.g., 90) and hitting **Tab** on the keyboard. The two computed values corresponding to the certainty value will then be displayed. In this example, there is a 5% probability that income will be below \$0.2781 and another 5% probability that income will be above \$1.3068; that is, the two-tailed confidence interval is a symmetrical interval centered on the median or 50th percentile value. Thus, both tails will have the same probability.

Alternatively, a one-tail probability can be computed. Figure I.13 shows a **Left-Tail** selection at 95% confidence (i.e., choose **Left-Tail** as the type, enter **95** as

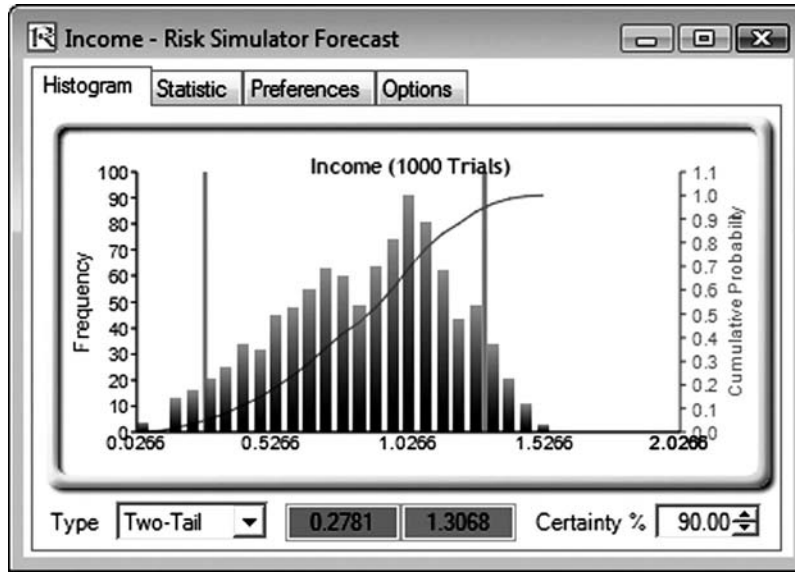


FIGURE I.12 Forecast chart two-tailed confidence interval

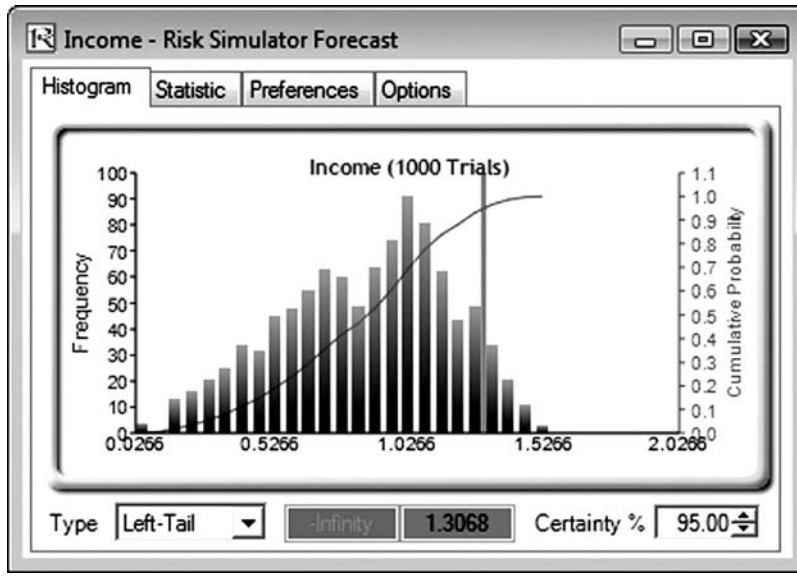


FIGURE I.13 Forecast chart one-tailed confidence interval

the certainty level, and hit **Tab** on the keyboard). This means that there is a 95% probability that the income will be below \$1.3068 (i.e., 95% on the left tail of \$1.3068) or a 5% probability that income will be above \$1.3068, corresponding perfectly with the results seen in Figure I.12.

In addition to evaluating the confidence interval (i.e., given a probability level and finding the relevant income values), you can determine the probability of a given income value (Figure I.14). For instance, what is the probability that income will be

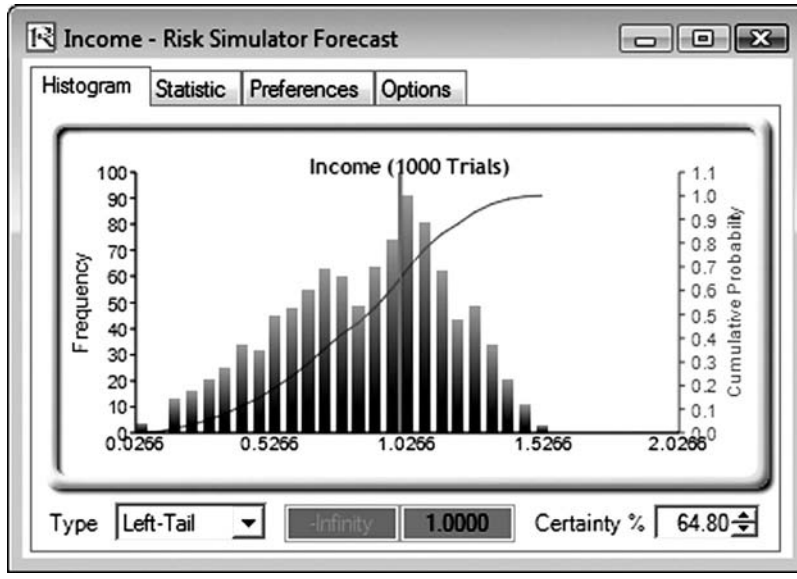


FIGURE I.14 Forecast chart left-tail probability evaluation

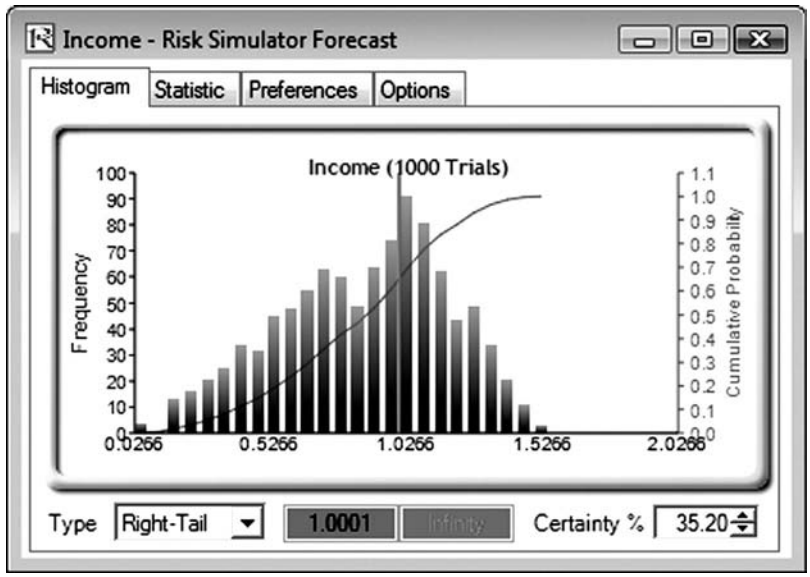


FIGURE I.15 Forecast chart right-tail probability evaluation

less than or equal to \$1? To do this, select the **Left-Tail** probability type, enter **1** into the value input box, and hit **Tab**. The corresponding certainty will be computed. (In this case, there is a 64.80% probability income will be at or below \$1.)

For the sake of completeness, you can select the **Right-Tail** probability type, enter the value *1* in the value input box, and hit **Tab** (Figure I.15). The resulting probability indicates the right-tail probability past the value 1; that is, the probability of income exceeding \$1. In this case, we see that there is a 35.20% probability of income at or exceeding \$1.

Note that the forecast window is resizable by clicking on and dragging the bottom right corner of the window.

Finally, it is always advisable that before rerunning a simulation, you reset the current simulation by selecting **Risk Simulator | Reset Simulation**.

CORRELATIONS AND PRECISION CONTROL

The correlation coefficient is a measure of the strength and direction of the relationship between two variables and can take on any values between –1.0 and +1.0; that is, the correlation coefficient can be decomposed into its direction or sign (positive or negative relationship between two variables) and the magnitude or strength of the relationship (the higher the absolute value of the correlation coefficient, the stronger the relationship).

The correlation coefficient can be computed in several ways. The first approach is to manually compute the correlation coefficient r of a pair of variables x and y using:

$$r_{x,y} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - \left(\sum x_i\right)^2}\sqrt{n\sum y_i^2 - \left(\sum y_i\right)^2}}$$

The second approach is to use Excel's *CORREL* function. For instance, if the 10 data points for x and y are listed in cells A1:B10, then the Excel function to use is *CORREL (A1:A10, B1:B10)*.

The third approach is to run Risk Simulator's *Multi-Variable Distributional Fitting Tool,* and the resulting correlation matrix will be computed and displayed.

It is important to note that correlation does not imply causation. Two completely unrelated random variables might display some correlation, but this does not imply any causation between the two (e.g., sunspot activity and events in the stock market are correlated, but there is no causation between the two).

There are two general types of correlations: parametric and nonparametric correlations. Pearson's correlation coefficient is the most common correlation measure and usually is referred to simply as the correlation coefficient. However, Pearson's correlation is a parametric measure, which means that it requires both correlated variables to have an underlying normal distribution and that the relationship between the variables is linear. When these conditions are violated, which is often the case in Monte Carlo simulations, the nonparametric counterparts become more important. Spearman's rank correlation and Kendall's tau are the two nonparametric alternatives. The Spearman correlation is used most commonly and is most appropriate when applied in the context of Monte Carlo simulation—there is no dependence on normal distributions or linearity, meaning that correlations between different variables with different distribution can be applied. In order to compute the Spearman correlation, first rank all the x and y variable values and then apply the Pearson's correlation computation.

Risk Simulator uses the more robust nonparametric Spearman's rank correlation. However, to simplify the simulation process and to be consistent with Excel's correlation function, the correlation user inputs required are the Pearson's correlation coefficient. Risk Simulator then applies its own algorithms to convert them into Spearman's rank correlation, thereby simplifying the process.

Applying Correlations in Risk Simulator

Correlations can be applied in Risk Simulator in several ways:

- When defining assumptions, simply enter the correlations into the correlation grid in the set input assumption dialog in Figure I.6.
- With existing data, run the *Multi-Variable Distribution Fitting Tool* to perform distributional fitting and to obtain the correlation matrix between pairwise variables. If a simulation profile exists, the assumptions fitted automatically will contain the relevant correlation values.
- With the use of a direct-input correlation matrix, click on **Risk Simulator | Edit Correlations** after multiple assumptions have been set, to view and edit the correlation matrix used in the simulation.

Note that the correlation matrix must be positive definite; that is, the correlation must be mathematically valid. For instance, suppose you are trying to correlate three variables: grades of graduate students in a particular year, the number of beers they

consume a week, and the number of hours they study a week. You would assume that these correlation relationships exist:

Grades and Beer:	–	The more they drink, the lower the grades (no show on exams)
Grades and Study:	+	The more they study, the higher the grades
Beer and Study:	–	The more they drink, the less they study (drunk and partying all the time)

However, if you input a negative correlation between Grades and Study and assuming that the correlation coefficients have high magnitudes, the correlation matrix will be nonpositive definite. It would defy logic, correlation requirements, and matrix mathematics. However, smaller coefficients sometimes still can work, even with the bad logic. When a nonpositive definite or bad correlation matrix is entered, Risk Simulator automatically informs you of the error and offers to adjust these correlations to something that is semipositive definite while still maintaining the overall structure of the correlation relationship (the same signs as well as the same relative strengths).

Effects of Correlations in Monte Carlo Simulation

Although the computations required to correlate variables in a simulation are complex, the resulting effects are fairly clear. Figure I.16 shows a simple correlation model (Correlation Risk Effects Model in the example folder). The calculation for revenue is simply price multiplied by quantity. The same model is replicated for no correlations, positive correlation (+0.9), and negative correlation (–0.9) between price and quantity.

The resulting statistics are shown in Figure I.17. Notice that the standard deviation of the model without correlations is 0.1450, compared to 0.1886 for the positive correlation model and 0.0717 for the negative correlation model. That is, for simple models with positive relationships (e.g., additions and multiplications), negative correlations tend to reduce the average spread of the distribution and create a tighter and more concentrated forecast distribution as compared to positive correlations with larger average spreads. However, the mean remains relatively stable. This implies that correlations do little to change the expected value of projects but can reduce or increase a project's risk. Recall in financial theory that

Correlation Model

	Without Correlation	Positive Correlation	Negative Correlation
Price	$2.00	$2.00	$2.00
Quantity	1.00	1.00	1.00
Revenue	$2.00	$2.00	$2.00

FIGURE I.16 Simple correlation model

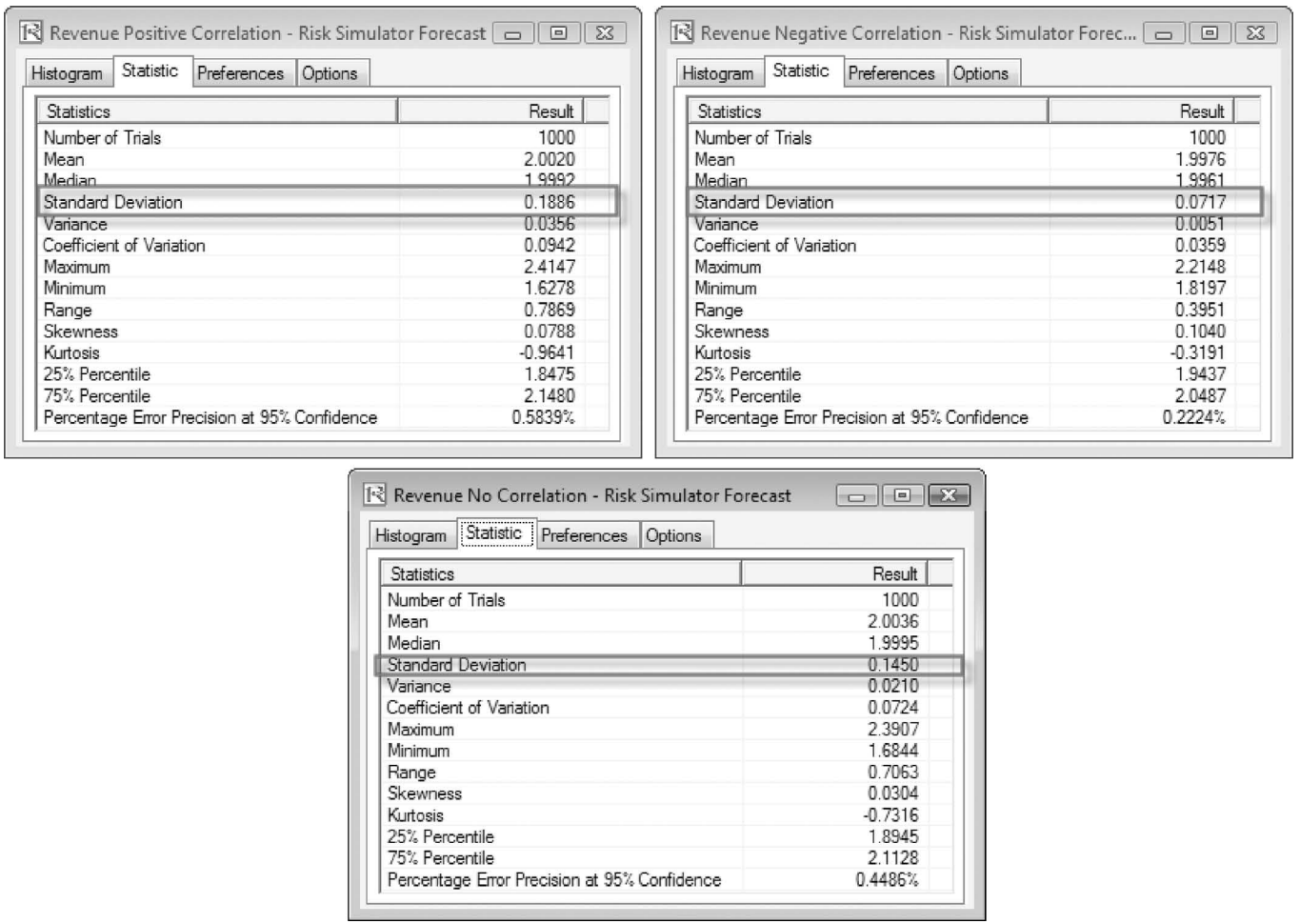


FIGURE I.17 Correlation results

Spearman's Nonlinear Rank Correlation on Raw Data Extracted from Simulation

Price Negative Correlation	Quantity Negative Correlation	Correlation	Price Positive Correlation	Quantity Positive Correlation	Correlation
676	145	**–0.90**	102	158	**0.89**
368	452		461	515	
264	880		515	477	
235	877		874	833	
122	711		769	792	
490	641		481	471	
336	638		627	446	
495	383		82	190	
241	568		659	674	
651	571		188	286	
854	59		458	439	
66	950		981	972	
707	262		528	569	
943	186		865	812	

FIGURE I.18 Correlations recovered

negatively correlated variables, projects, or assets when combined in a portfolio tend to create a diversification effect where the overall risk is reduced. Therefore, we see a smaller standard deviation for the negatively correlated model.

In a positively related model (e.g., $A + B = C$ or $A \times B = C$), a negative correlation reduces the risk (standard deviation and all other second moments of the distribution) of the result (C) whereas a positive correlation between the inputs (A and B) will increase the overall risk. The opposite is true for a negatively related model (e.g., $A - B = C$ or $A/B = C$), where a positive correlation between the inputs will reduce the risk and a negative correlation increases the risk. In more complex models, as is often the case in real-life situations, the effects will be unknown *a priori* and can be determined only after a simulation is run.

Figure I.18 illustrates the results after running a simulation, extracting the raw data of the assumptions, and computing the correlations between the variables. The figure shows that the input assumptions are recovered in the simulation; that is, you enter +0.9 and –0.9 correlations and the resulting simulated values have the same correlations. Clearly there will be minor differences from one simulation run to another, but when enough trials are run, the resulting recovered correlations approach those that were inputted.

TORNADO AND SENSITIVITY TOOLS IN SIMULATION

One of the powerful simulation tools in Risk Simulator is tornado analysis—it captures the static impacts of each variable on the outcome of the model; that is, the tool automatically perturbs each variable in the model a preset amount, captures the fluctuation on the model's forecast or final result, and lists the resulting perturbations ranked from the most significant to the least. Figures I.19 through I.24 illustrate the application of a tornado analysis. For instance, Figure I.19 is a sample discounted

Discounted Cash Flow Model

Base Year	*2005*	*Sum PV Net Benefits*	***$1,896.63***
Market Risk-Adjusted Discount Rate	*15.00%*	*Sum PV Investments*	***$1,800.00***
Private-Risk Discount Rate	*5.00%*	*Net Present Value*	***$96.63***
Annualized Sales Growth Rate	*2.00%*	*Internal Rate of Return*	*18.80%*
Price Erosion Rate	*5.00%*	*Return on Investment*	*5.37%*
Effective Tax Rate	*40.00%*		

	2005	2006	2007	2008	2009
Product A Avg Price/Unit	$10.00	$9.50	$9.03	$8.57	$8.15
Product B Avg Price/Unit	$12.25	$11.64	$11.06	$10.50	$9.98
Product C Avg Price/Unit	$15.15	$14.39	$13.67	$12.99	$12.34
Product A Sale Quantity ('000s)	50.00	51.00	52.02	53.06	54.12
Product B Sale Quantity ('000s)	35.00	35.70	36.41	37.14	37.89
Product C Sale Quantity ('000s)	20.00	20.40	20.81	21.22	21.65
Total Revenues	***$1,231.75***	***$1,193.57***	***$1,156.57***	***$1,120.71***	***$1,085.97***
Direct Cost of Goods Sold	$184.76	$179.03	$173.48	$168.11	$162.90
Gross Profit	***$1,046.99***	***$1,014.53***	***$983.08***	***$952.60***	***$923.07***
Operating Expenses	$157.50	$160.65	$163.86	$167.14	$170.48
Sales, General and Admin. Costs	$15.75	$16.07	$16.39	$16.71	$17.05
Operating Income (EBITDA)	***$873.74***	***$837.82***	***$802.83***	***$768.75***	***$735.54***
Depreciation	$10.00	$10.00	$10.00	$10.00	$10.00
Amortization	$3.00	$3.00	$3.00	$3.00	$3.00
EBIT	***$860.74***	***$824.82***	***$789.83***	***$755.75***	***$722.54***
Interest Payments	$2.00	$2.00	$2.00	$2.00	$2.00
EBT	***$858.74***	***$822.82***	***$787.83***	***$753.75***	***$720.54***
Taxes	$343.50	$329.13	$315.13	$301.50	$288.22
Net Income	***$515.24***	***$493.69***	***$472.70***	***$452.25***	***$432.33***
Noncash: Depreciation Amortization	$13.00	$13.00	$13.00	$13.00	$13.00
Noncash: Change in Net Working Capital	$0.00	$0.00	$0.00	$0.00	$0.00
Noncash: Capital Expenditures	$0.00	$0.00	$0.00	$0.00	$0.00
Free Cash Flow	***$528.24***	***$506.69***	***$485.70***	***$465.25***	***$445.33***
Investment Outlay	***$1,800.00***				
Financial Analysis					
Present Value of Free Cash Flow	***$528.24***	***$440.60***	***$367.26***	***$305.91***	***$254.62***
Present Value of Investment Outlay	***$1,800.00***	***$0.00***	***$0.00***	***$0.00***	***$0.00***
Net Cash Flows	($1,271.76)	$506.69	$485.70	$465.25	$445.33

FIGURE I.19 Sample discounted cash flow model

cash flow model where the input assumptions in the model are shown. The question is: What are the critical success drivers that affect the model's output the most? That is, what really drives the net present value of $96.63 or which input variable impacts this value the most?

The tornado chart tool can be obtained through **Risk Simulator | Tools | Tornado Analysis**. To follow along the first example, open the *Tornado and Sensitivity Charts (Linear)* file in the examples folder. Figure I.20 shows this sample model, where cell G6 containing the net present value is chosen as the target result to be analyzed. The target cell's precedents in the model are used in creating the tornado chart. Precedents are all the input variables that affect the outcome of the model. For instance, if the

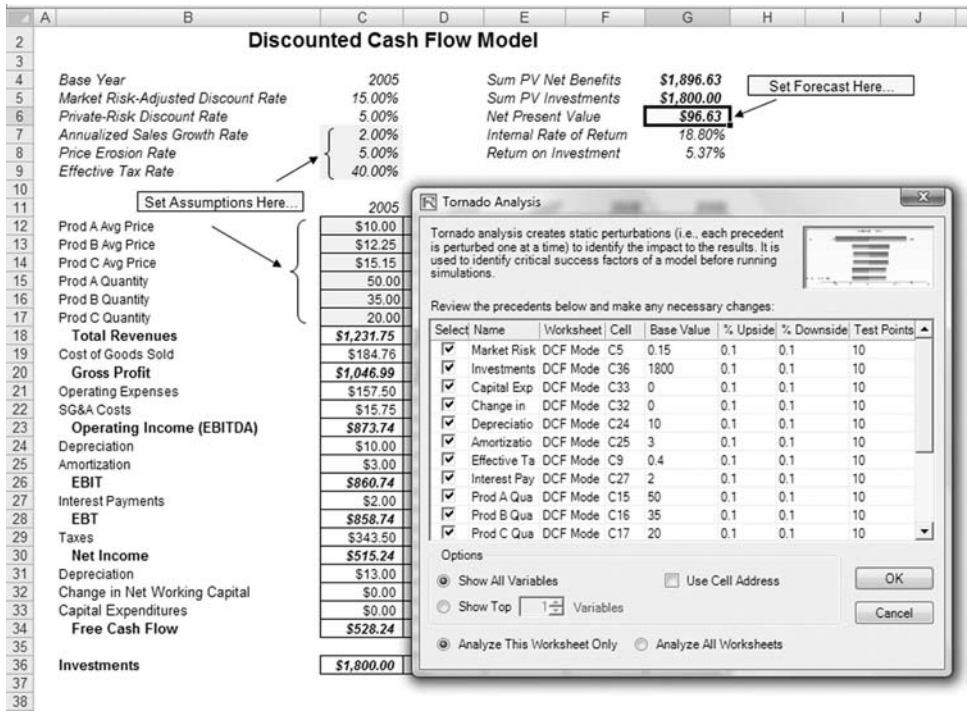

Discounted Cash Flow Model
Base Year 2005
Market Risk-Adjusted Discount Rate 15.00%
Private-Risk Discount Rate 5.00%
Annualized Sales Growth Rate 2.00%
Price Erosion Rate 5.00%
Effective Tax Rate 40.00%
Sum PV Net Benefits $1,896.63
Sum PV Investments $1,800.00
Net Present Value $96.63
Internal Rate of Return 18.80%
Return on Investment 5.37%
Set Forecast Here...
Set Assumptions Here...
Prod A Avg Price $10.00
Prod B Avg Price $12.25
Prod C Avg Price $15.15
Prod A Quantity 50.00
Prod B Quantity 35.00
Prod C Quantity 20.00
Total Revenues $1,231.75
Cost of Goods Sold $184.76
Gross Profit $1,046.99
Operating Expenses $157.50
SG&A Costs $15.75
Operating Income (EBITDA) $873.74
Depreciation $10.00
Amortization $3.00
EBIT $860.74
Interest Payments $2.00
EBT $858.74
Taxes $343.50
Net Income $515.24
Depreciation $13.00
Change in Net Working Capital $0.00
Capital Expenditures $0.00
Free Cash Flow $528.24
Investments $1,800.00
Tornado Analysis
Tornado analysis creates static perturbations (i.e., each precedent is perturbed one at a time) to identify the impact to the results. It is used to identify critical success factors of a model before running simulations.
Review the precedents below and make any necessary changes:
Select Name Worksheet Cell Base Value % Upside % Downside Test Points
Market Risk DCF Mode C5 0.15 0.1 0.1 10
Investments DCF Mode C36 1800 0.1 0.1 10
Capital Exp DCF Mode C33 0 0.1 0.1 10
Change in DCF Mode C32 0 0.1 0.1 10
Depreciatio DCF Mode C24 10 0.1 0.1 10
Amortizatio DCF Mode C25 3 0.1 0.1 10
Effective Ta DCF Mode C9 0.4 0.1 0.1 10
Interest Pay DCF Mode C27 2 0.1 0.1 10
Prod A Qua DCF Mode C15 50 0.1 0.1 10
Prod B Qua DCF Mode C16 35 0.1 0.1 10
Prod C Qua DCF Mode C17 20 0.1 0.1 10
Options
Show All Variables
Use Cell Address
Show Top 1 Variables
Analyze This Worksheet Only
Analyze All Worksheets
OK
Cancel

FIGURE I.20 Running a tornado analysis

model consists of $A = B + C$, and where $C = D + E$, then B, D, and E are the precedents for A (C is not a precedent as it is only an intermediate calculated value). Figure I.20 shows the testing range of each precedent variable used to estimate the target result. If the precedent variables are simple inputs, then the testing range will be a simple perturbation based on the range chosen (e.g., the default is $\pm 10\%$). Each precedent variable can be perturbed at different percentages if required. A wider range is important as it is better able to test extreme values rather than smaller perturbations around the expected values. In certain circumstances, extreme values may have a larger, smaller, or unbalanced impact (e.g., nonlinearities may occur where increasing or decreasing economies of scale and scope creep in for larger or smaller values of a variable) and only a wider range will capture this nonlinear impact.

PROCEDURE

Use these three steps to create a tornado analysis:

1. Select the single output cell (i.e., a cell with a function or equation) in an Excel model (e.g., cell G6 is selected in our example).
2. Select **Risk Simulator | Tools | Tornado Analysis**.

3. Review the precedents and rename them as appropriate (renaming the precedents to shorter names allows a more visually pleasing tornado and spider chart) and click **OK**. Alternatively, click on **Use Cell Address** to apply cell locations as the variable names.

Results Interpretation

Figure I.21 shows the resulting tornado analysis report, which indicates that capital investment has the largest impact on net present value (NPV), followed by tax rate, average sale price and quantity demanded of the product lines, and so forth. The report contains four distinct elements:

1. A statistical summary listing the procedure performed.
2. A sensitivity table (Figure I.22) shows the starting NPV base value of $96.63 and how each input is changed (e.g., Investment is changed from $1,800 to $1,980 on the upside with a +10% swing and from $1,800 to $1,620 on the downside with a –10% swing). The resulting upside and downside values on NPV are –$83.37 and $276.63, with a total change of $360, making it the variable with the highest impact on NPV. The precedent variables are ranked from the highest impact to the lowest impact.
3. The spider chart (Figure I.23) illustrates these effects graphically. The *y*-axis is the NPV target value while the *x*-axis depicts the percentage change on each of the precedent value. The central point is the base case value at $96.63 at 0% change from the base value of each precedent. Positively sloped lines indicate a positive relationship or effect; negatively sloped lines indicate a negative relationship (e.g., investment is negatively sloped, which means that the higher the investment level, the lower the NPV). The absolute value of the slope indicates the magnitude of the effect computed as the percentage change in the result given a percentage change in the precedent. A steep line indicates a higher impact on the NPV *y*-axis given a change in the precedent *x*-axis.
4. The tornado chart (Figure I.24) illustrates the results in another graphical manner, where the highest-impacting precedent is listed first. The *x*-axis is the NPV value, with the center of the chart being the base case condition. Green (lighter) bars in the chart indicate a positive effect; red (darker) bars indicate a negative effect. Therefore, for investments, the red (darker) bar on the right side indicate a negative effect of investment on higher NPV—in other words, capital investment and NPV are negatively correlated. The opposite is true for price and quantity of products A to C (their green or lighter bars are on the right side of the chart).

Notes

Remember that tornado analysis is a *static* sensitivity analysis applied on each input variable in the model—that is, each variable is perturbed individually, and the resulting effects are tabulated. This makes tornado analysis a key component to execute before running a simulation. Capturing and identifying the most important impact drivers in the model is one of the very first steps in risk analysis. The next step

Tornado and Spider Charts

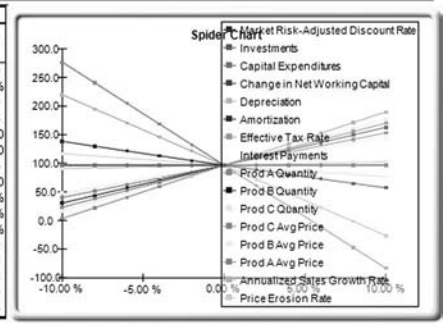
Statistical Summary

One of the powerful simulation tools is the tornado chart—it captures the static impacts of each variable on the outcome of the model. That is, the tool automatically perturbs each precedent variable in the model a user-specified preset amount, captures the fluctuation on the model's forecast or final result, and lists the resulting perturbations ranked from the most significant to the least. Precedents are all the input and intermediate variables that affect the outcome of the model. For instance, if the model consists of A = B + C, where C = D + E, then B, D, and E are the precedents for A (C is not a precedent as it is only an intermediate calculated value). The range and number of values perturbed is user-specified and can be set to test extreme values rather than smaller perturbations around the expected values. In certain circumstances, extreme values may have a larger, smaller, or unbalanced impact (e.g., nonlinearities may occur where increasing or decreasing economies of scale and scope creep occurs for larger or smaller values of a variable) and only a wider range will capture this nonlinear impact.

A tornado chart lists all the inputs that drive the model, starting from the input variable that has the most effect on the results. The chart is obtained by perturbing each precedent input at some consistent range (e.g., ±10% from the base case) one at a time, and comparing their results to the base case. A spider chart looks like a spider with a central body and its many legs protruding. The positively sloped line indicates a positive relationship, while a negatively sloped line indicates a negative relationship. Further, spider charts can be used to visualize linear and nonlinear relationships. The tornado and spider charts help identify the critical success factors of an output cell in order to identify the inputs to simulate. The identified critical variables that are uncertain are the ones that should be simulated. Do not waste time simulating variables that are neither uncertain nor have little impact on the results.

Result

	Base Value: 96.6261638553219			Input Changes		
Precedent Cell	Output Downside	Output Upside	Effective Range	Input Downside	Input Upside	Base Case Value
Investments	276.6262	-83.3738	360.00	$1,620.00	$1,980.00	$1,800.00
Effective Tax Rate	219.7269	-26.4746	246.20	36.00%	44.00%	40.00%
Prod A Avg Price	3.425542	189.8268	186.40	$9.00	$11.00	$10.00
Prod B Avg Price	16.70663	176.5457	159.84	$11.03	$13.48	$12.25
Prod A Quantity	23.1775	170.0748	146.90	45.00	55.00	50.00
Prod B Quantity	30.533	162.7193	132.19	31.50	38.50	35.00
Prod C Avg Price	40.14659	153.1057	112.96	$13.64	$16.67	$15.15
Prod C Quantity	48.04737	145.205	97.16	18.00	22.00	20.00
Market Risk-Adjusted I	138.2391	57.02984	81.21	13.50%	16.50%	15.00%
Price Erosion Rate	116.8038	76.64095	40.16	4.50%	5.50%	5.00%
Annualized Sales Grov	90.58835	102.6854	12.10	1.80%	2.20%	2.00%
Depreciation	95.08417	98.16816	3.08	$9.00	$11.00	$10.00
Amortization	96.16357	97.08876	0.93	$2.70	$3.30	$3.00
Interest Payments	97.08876	96.16357	0.93	$1.80	$2.20	$2.00
Change in Net Workin	96.62616	96.62616	0.00	$0.00	$0.00	$0.00
Capital Expenditures	96.62616	96.62616	0.00	$0.00	$0.00	$0.00



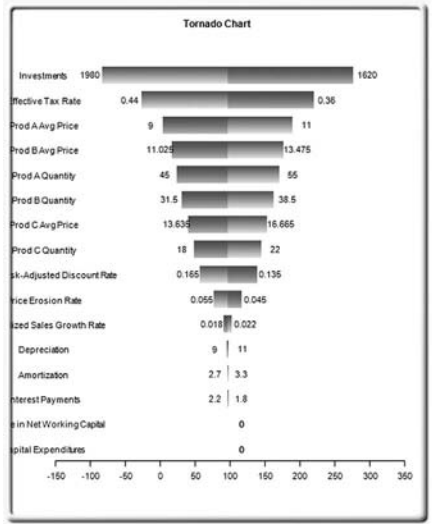


FIGURE I.21 Tornado analysis report

is to identify which of these important impact drivers are uncertain. These uncertain impact drivers are the critical success drivers of a project; the results of the model depend on these critical success drivers. These variables are the ones that should be simulated. Do not waste time simulating variables that are neither uncertain nor have little impact on the results. Tornado charts assist in identifying these critical success drivers quickly and easily. Following this example, it might be that price and quantity should be simulated, assuming if the required investment and effective tax rate are both known in advance and unchanging.

	Base Value: 96.6261638553219			Input Changes		
Precedent Cell	*Output Downside*	*Output Upside*	*Effective Range*	*Input Downside*	*Input Upside*	*Base Case Value*
Investment	$276.63	($83.37)	360.00	$1,620.00	$1,980.00	$1,800.00
Tax Rate	$219.73	($26.47)	246.20	36.00%	44.00%	40.00%
A Price	$3.43	$189.83	186.40	$9.00	$11.00	$10.00
B Price	$16.71	$176.55	159.84	$11.03	$13.48	$12.25
A Quantity	$23.18	$170.07	146.90	45.00	55.00	50.00
B Quantity	$30.53	$162.72	132.19	31.50	38.50	35.00
C Price	$40.15	$153.11	112.96	$13.64	$16.67	$15.15
C Quantity	$48.05	$145.20	97.16	18.00	22.00	20.00
Discount Rate	$138.24	$57.03	81.21	13.50%	16.50%	15.00%
Price Erosion	$116.80	$76.64	40.16	4.50%	5.50%	5.00%
Sales Growth	$90.59	$102.69	12.10	1.80%	2.20%	2.00%
Depreciation	$95.08	$98.17	3.08	$9.00	$11.00	$10.00
Interest	$97.09	$96.16	0.93	$1.80	$2.20	$2.00
Amortization	$96.16	$97.09	0.93	$2.70	$3.30	$3.00
Capex	$96.63	$96.63	0.00	$0.00	$0.00	$0.00
Net Capital	$96.63	$96.63	0.00	$0.00	$0.00	$0.00

FIGURE I.22 Sensitivity table

Although the tornado chart is easier to read, the spider chart is important to determine if there are any nonlinearities in the model. For instance, Figure I.25 shows another spider chart where nonlinearities are fairly evident (the lines on the graph are not straight but curved). The example model used is **Tornado and Sensitivity Charts (Nonlinear)**, which applies the Black-Scholes option pricing model. Such nonlinearities cannot be easily ascertained from a tornado chart and may be important information in the model or may provide decision makers important insight

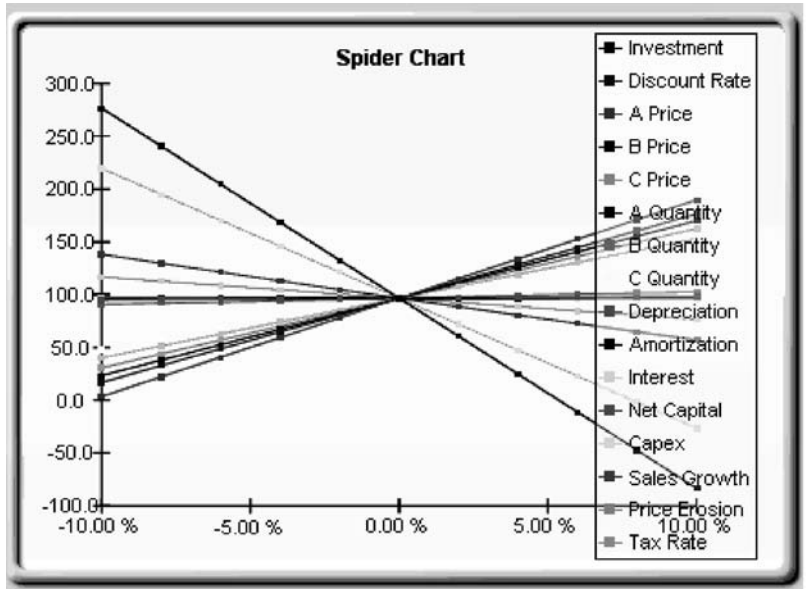


FIGURE I.23 Spider chart

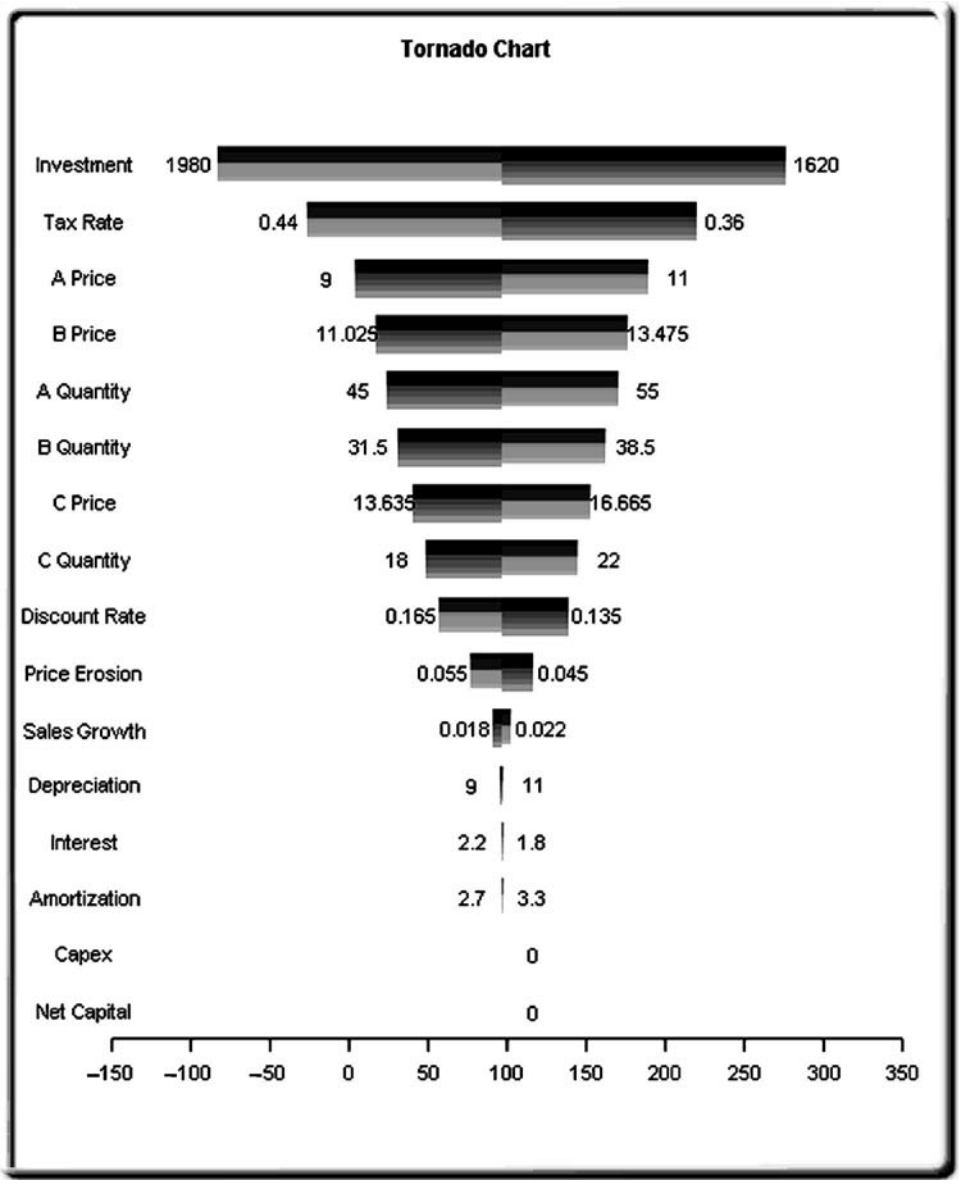


FIGURE 1.24 Tornado chart

into the model's dynamics. For instance, in this Black-Scholes model, the fact that stock price and strike price are nonlinearly related to the option value is important to know. This characteristic implies that option value will not increase or decrease proportionally to the changes in stock or strike price and that there might be some interactions between these two prices as well as other variables. As another example, an engineering model depicting nonlinearities might indicate that a particular part or component, when subjected to a high enough force or tension, will break. Clearly, it is important to understand such nonlinearities.

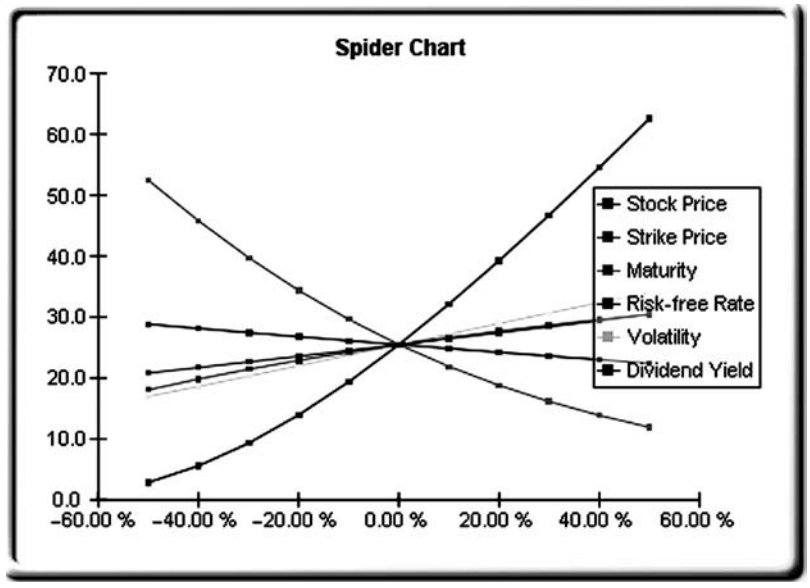


FIGURE I.25 Nonlinear spider chart

SENSITIVITY ANALYSIS

A related feature is sensitivity analysis. While tornado analysis (tornado charts and spider charts) applies static perturbations *before* a simulation run, sensitivity analysis applies dynamic perturbations created *after* the simulation run. Tornado and spider charts are the results of static perturbations, meaning that each precedent or assumption variable is perturbed a preset amount one at a time, and the fluctuations in the results are tabulated. In contrast, sensitivity charts are the results of dynamic perturbations in the sense that multiple assumptions are perturbed simultaneously and their interactions in the model and correlations among variables are captured in the fluctuations of the results. Tornado charts therefore identify which variables drive the results the most and hence are suitable for simulation; sensitivity charts identify the impact to the results when multiple interacting variables are simulated together in the model. This effect is clearly illustrated in Figure I.26. Notice that the ranking of critical success drivers is similar to the tornado chart in the previous examples. However, if correlations are added between the assumptions, Figure I.27 shows a very different picture. Notice, for instance, price erosion had little impact on NPV, but when some of the input assumptions are correlated, the interaction that exists between these correlated variables makes price erosion have more impact. Note that tornado analysis cannot capture these correlated dynamic relationships. Only after a simulation is run will such relationships become evident in a sensitivity analysis. A tornado chart's presimulation critical success factors therefore sometimes will be different from a sensitivity chart's postsimulation critical success factors. The postsimulation critical success factors should be the ones that are of interest, as these more readily capture the interactions of the model precedents.

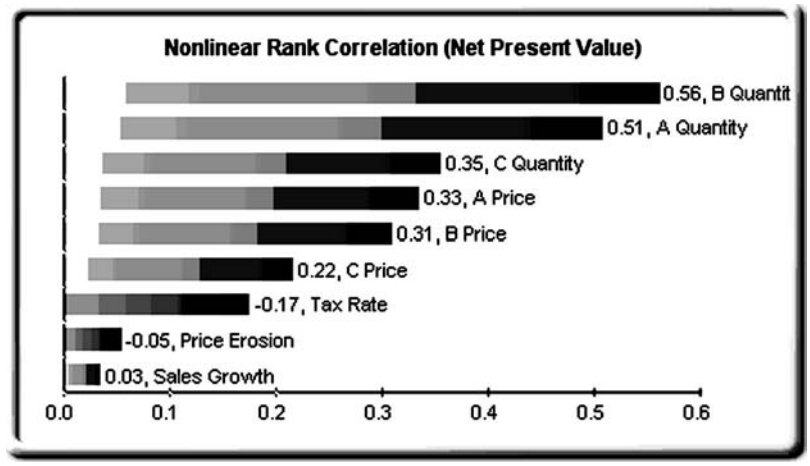


FIGURE I.26 Sensitivity chart without correlations

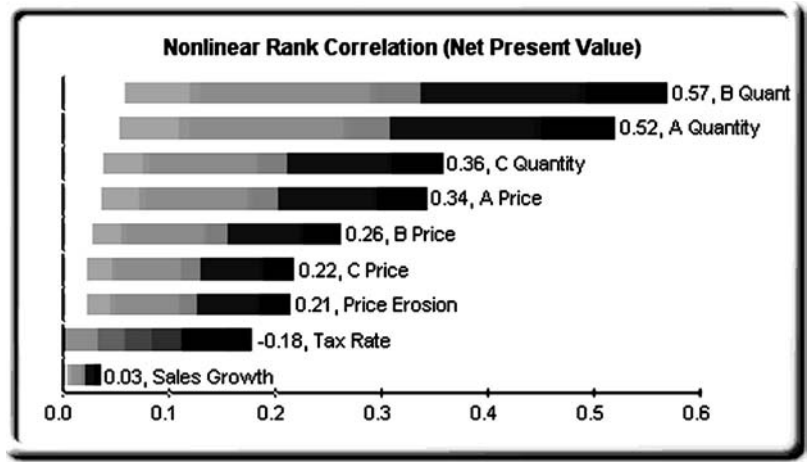


FIGURE I.27 Sensitivity chart with correlations

PROCEDURE

Use these three steps to create a sensitivity analysis:

1. Open or create a model, define assumptions and forecasts, and run the simulation—the example here uses the *Tornado and Sensitivity Charts (Linear)* file.
2. Select **Risk Simulator | Tools | Sensitivity Analysis.**
3. Select the forecast of choice to analyze and click **OK** (Figure I.28).

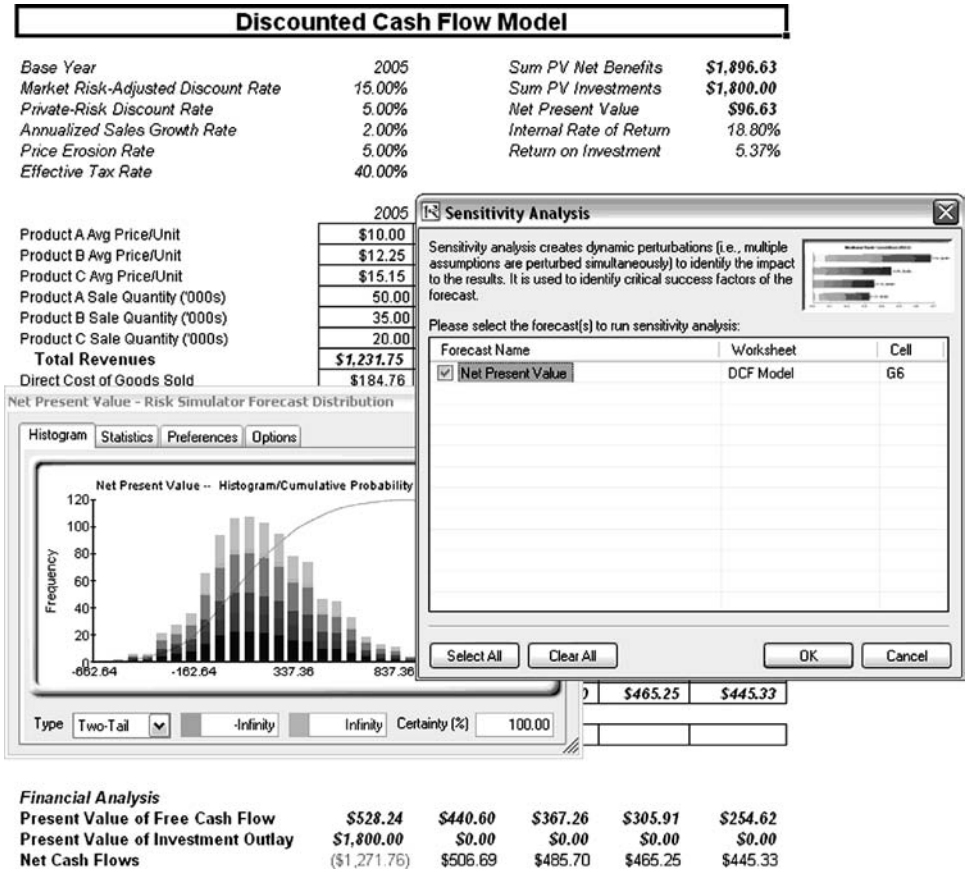


FIGURE I.28 Running sensitivity analysis

Note that sensitivity analysis cannot be run unless assumptions and forecasts have been defined and a simulation has been run.

Results Interpretation

The results of the sensitivity analysis comprise a report and two key charts. The first is a nonlinear rank correlation chart (Figure I.29) that ranks from highest to lowest the assumption-forecast correlation pairs. These correlations are nonlinear and nonparametric, making them free of any distributional requirements (i.e., an assumption with a Weibull distribution can be compared to another with a beta distribution). The results from this chart are fairly similar to that of the tornado analysis seen previously (of course without the capital investment value, which we decided was a known value and hence was not simulated), with one special exception. Tax rate was relegated to a much lower position in the sensitivity analysis chart (Figure I.29) as compared to the tornado chart (Figure I.24). This is because by itself, tax rate will have a significant impact. Once the other variables are interacting in the model, however, it appears that tax rate has less of a dominant effect. This is because

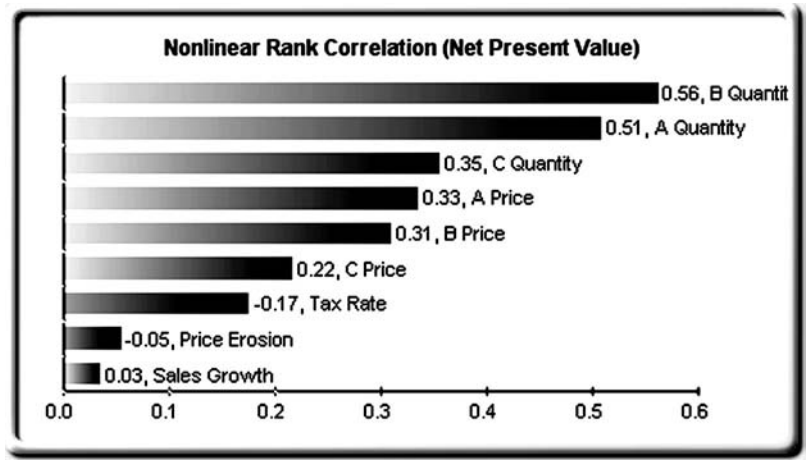


FIGURE I.29 Rank correlation chart

tax rate has a smaller distribution, as historical tax rates tend not to fluctuate too much, and also because tax rate is a straight percentage value of the income before taxes, whereas other precedent variables have a larger effect on NPV. This example proves that it is important to perform sensitivity analysis after a simulation run to ascertain if there are any interactions in the model and if the effects of certain variables still hold. The second chart (Figure I.30) illustrates the percent variation explained; that is, of the fluctuations in the forecast, how much of the variation can be explained by each of the assumptions after accounting for all the interactions among variables? Notice that the sum of all variations explained is usually close to 100% (sometimes other elements impact the model, but they cannot be captured

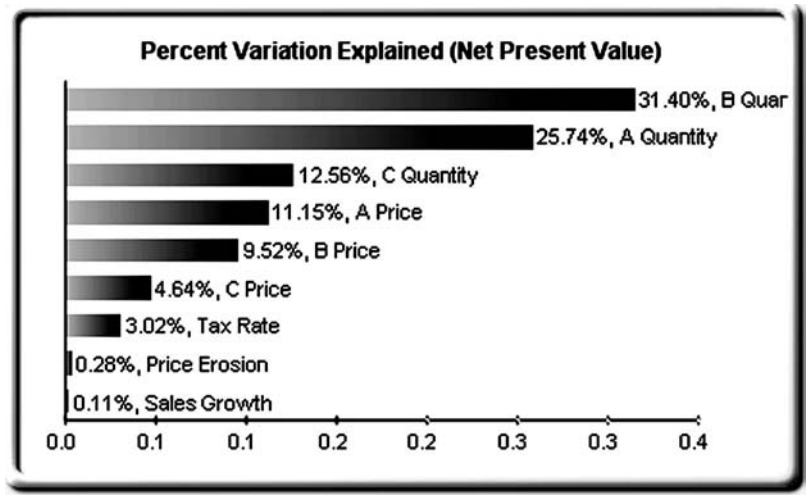


FIGURE I.30 Contribution to variance chart

here directly), and if correlations exist, the sum may sometimes exceed 100% (due to the interaction effects that are cumulative).

Notes

Tornado analysis is performed before a simulation run while sensitivity analysis is performed after a simulation run. Spider charts in tornado analysis can consider nonlinearities while rank correlation charts in sensitivity analysis can account for nonlinear and distributional-free conditions.

DISTRIBUTIONAL FITTING: SINGLE VARIABLE AND MULTIPLE VARIABLES

Another powerful simulation tool is distributional fitting; that is, which distribution does an analyst or engineer use for a particular input variable in a model? What are the relevant distributional parameters? If no historical data exist, then the analyst must make assumptions about the variables in question. One approach is to use the Delphi method, where a group of experts are tasked with estimating the behavior of each variable. For instance, a group of mechanical engineers can be tasked with evaluating the extreme possibilities of the diameter of a spring coil through rigorous experimentation or guesstimates. These values can be used as the variable's input parameters (e.g., uniform distribution with extreme values between 0.5 and 1.2). When testing is not possible (e.g., market share and revenue growth rate), management still can make estimates of potential outcomes and provide the best-case, most-likely case, and worst-case scenarios, whereupon a triangular or custom distribution can be created.

However, if reliable historical data are available, distributional fitting can be accomplished. Assuming that historical patterns hold and that history tends to repeat itself, historical data can be used to find the best-fitting distribution with their relevant parameters to better define the variables to be simulated. Clearly, adjustments to the forecast value can be made (e.g., structural shifts and adjustments) as required, to reflect future expectations. Figures I.31 through I.33 illustrate a distributional-fitting example. The next discussion uses the *Data Fitting* file in the examples folder.

PROCEDURE

Use these four steps to perform a distributional-fitting model:

1. Open a spreadsheet with existing data for fitting (e.g., use the *Data Fitting* example file from the **Risk Simulator | Example Models** menu).
2. Select the data you wish to fit, not including the variable name. (Data should be in a single column with multiple rows.)
3. Select **Risk Simulator | Tools | Distributional Fitting (Single-Variable)**. Decide if you wish to fit to continuous or discrete distributions.

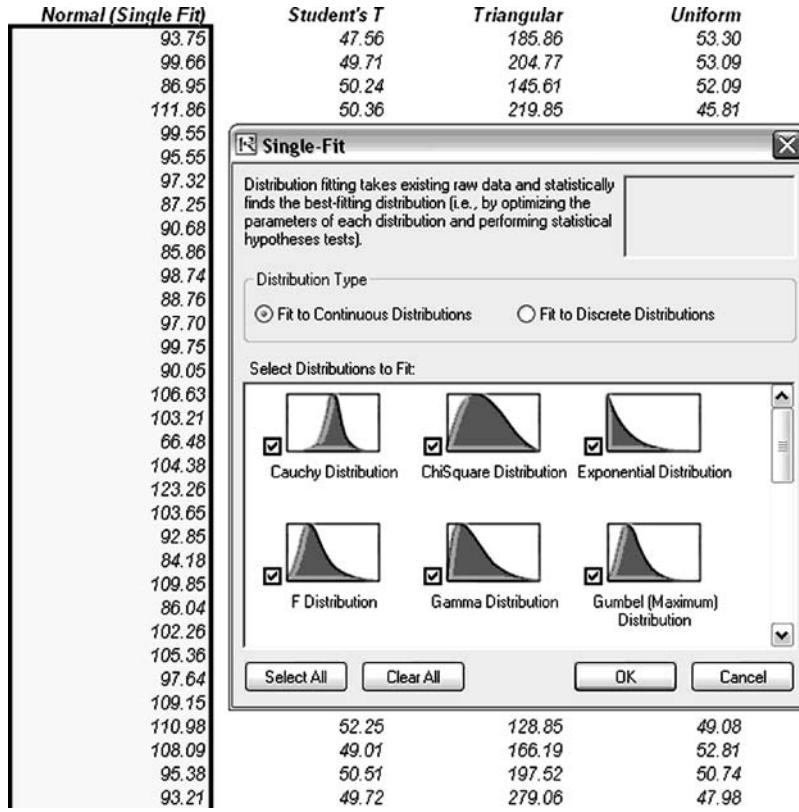


FIGURE I.31 Single-variable distributional fitting

4. Select the specific distributions you wish to fit to or keep the default where all distributions are selected and click **OK** (Figure I.31).
5. Review the results of the fit, choose the relevant distribution you want, and click **OK** (Figure I.32).

Results Interpretation

The null hypothesis (H_o) being tested is such that the fitted distribution is the same distribution as the population from which the sample data to be fitted comes. Thus, if the computed *p*-value is lower than a critical alpha level (typically 0.10 or 0.05), then the distribution is the wrong distribution. Conversely, the *higher the* p*-value, the better the distribution fits the data*. Roughly, you can think of *p*-value as a *percentage explained*; that is, if the *p*-value is 1.00 (Figure I.32), then setting a normal distribution with a mean of 100.67 and a standard deviation of 10.40 explains close to 100% of the variation in the data, indicating an especially good fit. The data was from a 1,000-trial simulation in Risk Simulator based on a normal distribution with a mean of 100 and a standard deviation of 10. Because only 1,000 trials were simulated, the resulting distribution is fairly close to the specified distributional parameters, and in this case, about a 100% precision.

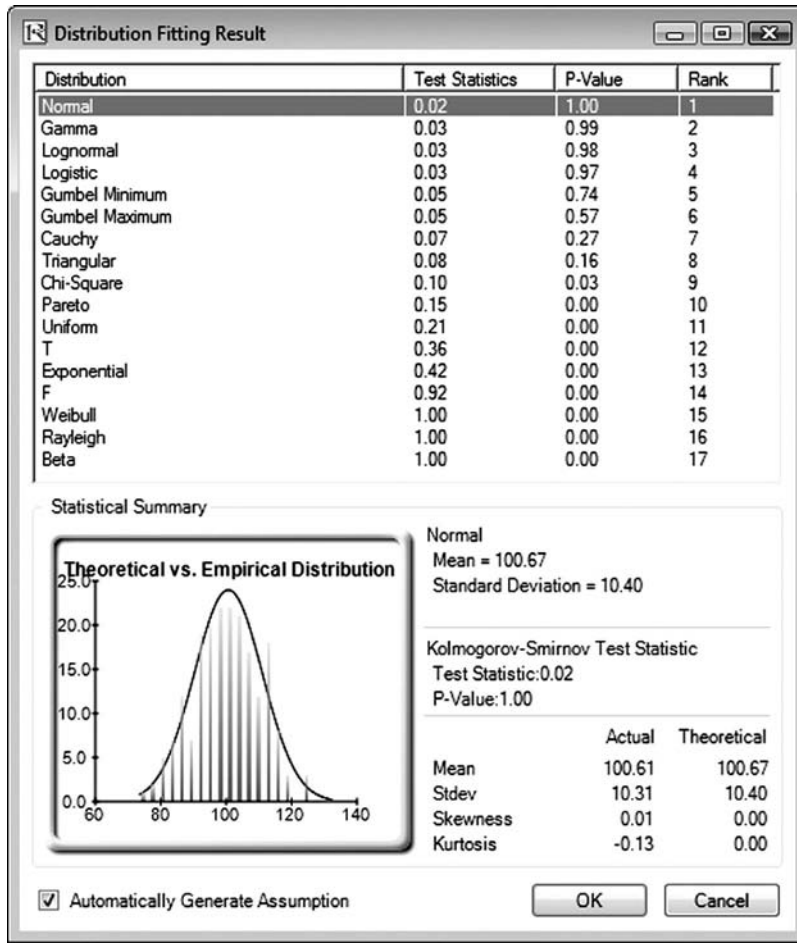


FIGURE I.32 Distributional fitting result

Both the results (Figure I.32) and the report (Figure I.33) show the test statistic, *p*-value, theoretical statistics (based on the selected distribution), empirical statistics (based on the raw data), the original data (to maintain a record of the data used), and the assumption complete with the relevant distributional parameters (i.e., if you selected the option to automatically generate assumption and if a simulation profile already exists). The results also rank all the selected distributions and how well they fit the data.

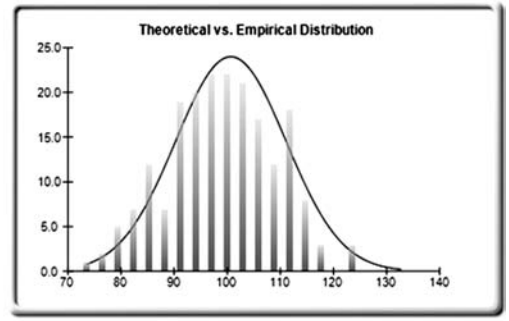
BOOTSTRAP SIMULATION

Bootstrap simulation is a simple technique that estimates the reliability or accuracy of forecast statistics or other sample raw data. Bootstrap simulation can be used to answer a lot of confidence and precision-based questions in simulation. For instance,

Statistical Summary

Fitted Assumption	100.61
Fitted Distribution	**Normal**
Mean	100.67
Sigma	10.40
Kolmogorov-Smirnov Statistic	0.02
P-Value for Test Statistic	0.9996

	Actual	Theoretical
Mean	100.61	100.67
Standard Deviation	10.31	10.40
Skewness	0.01	0.00
Excess Kurtosis	-0.13	0.00



Original Fitted Data

73.53	78.21	78.52	79.50	79.72	79.74	81.56	82.08	82.68	82.75	83.34	83.64	84.09
84.66	85.00	85.35	85.51	86.04	86.79	86.82	86.91	87.02	87.03	87.45	87.53	87.66
88.05	88.45	88.51	89.95	90.19	90.54	90.68	90.96	91.25	91.49	91.56	91.94	92.06
92.36	92.41	92.45	92.70	92.80	92.84	93.21	93.26	93.48	93.73	93.75	93.77	93.82
94.00	94.15	94.51	94.57	94.64	94.69	94.95	95.57	95.62	95.71	95.78	95.83	95.97
96.20	96.24	96.40	96.43	96.47	96.81	96.88	97.00	97.07	97.21	97.23	97.48	97.70
97.77	97.85	98.15	98.17	98.24	98.28	98.32	98.33	98.35	98.65	99.03	99.27	99.46
99.47	99.55	99.73	99.96	100.08	100.24	100.36	100.42	100.44	100.48	100.49	100.83	101.17
101.28	101.34	101.45	101.46	101.55	101.73	101.74	101.81	102.29	102.55	102.58	102.60	102.70
103.17	103.21	103.22	103.32	103.34	103.45	103.65	103.66	103.72	103.81	103.90	103.99	104.46
104.57	104.76	105.20	105.44	105.50	105.52	105.58	105.66	105.87	105.90	105.90	106.29	106.35
106.59	107.01	107.68	107.70	107.93	108.17	108.20	108.34	108.42	108.43	108.49	108.70	109.15
109.22	109.35	109.52	109.75	110.04	110.16	110.25	110.54	111.05	111.06	111.44	111.76	111.90
111.95	112.07	112.19	112.29	112.32	112.42	112.48	112.85	112.92	113.50	113.59	113.63	113.70
114.13	114.14	114.21	114.91	114.95	115.40	115.58	115.66	116.58	116.98	117.60	118.67	119.24
119.52	124.14	124.16	124.39	132.30								

FIGURE I.33 Distributional fitting report

suppose an identical model (with identical assumptions and forecasts but without any random seeds) is run by 100 different people; the results will clearly be slightly different. The question is, if we collected all the statistics from these 100 people, how will the mean be distributed, or the median, or the skewness, or excess kurtosis? Suppose one person has a mean value of, say, 1.50 while another has 1.52. Are these two values statistically significantly different from one another, or are they statistically similar and the slight difference is due entirely to random chance? What about 1.53? So, how far is far enough to say that the values are statistically different? In addition, if a model's resulting skewness is –0.19, is this forecast distribution negatively skewed or is it statistically close enough to zero to state that this distribution is symmetrical and not skewed? Thus, if we bootstrapped this forecast 100 times (i.e., run a 1,000-trial simulation for 100 times and collect the 100 skewness coefficients), the skewness distribution would indicate how far zero is away from –0.19. If the 90% confidence on the bootstrapped skewness distribution contains the value zero, then we can state that on a 90% confidence level, this distribution is symmetrical and not skewed, and the value –0.19 is statistically close enough to zero. Otherwise, if zero falls outside of this 90% confidence area, then this distribution is negatively skewed. The same analysis can be applied to excess kurtosis and other statistics.

Essentially, bootstrap simulation is a hypothesis-testing tool. Classical methods used in the past relied on mathematical formulas to describe the accuracy of sample statistics. These methods assume that the distribution of a sample statistic approaches a normal distribution, making the calculation of the statistic's standard error or confidence interval relatively easy. However, when a statistic's sampling distribution

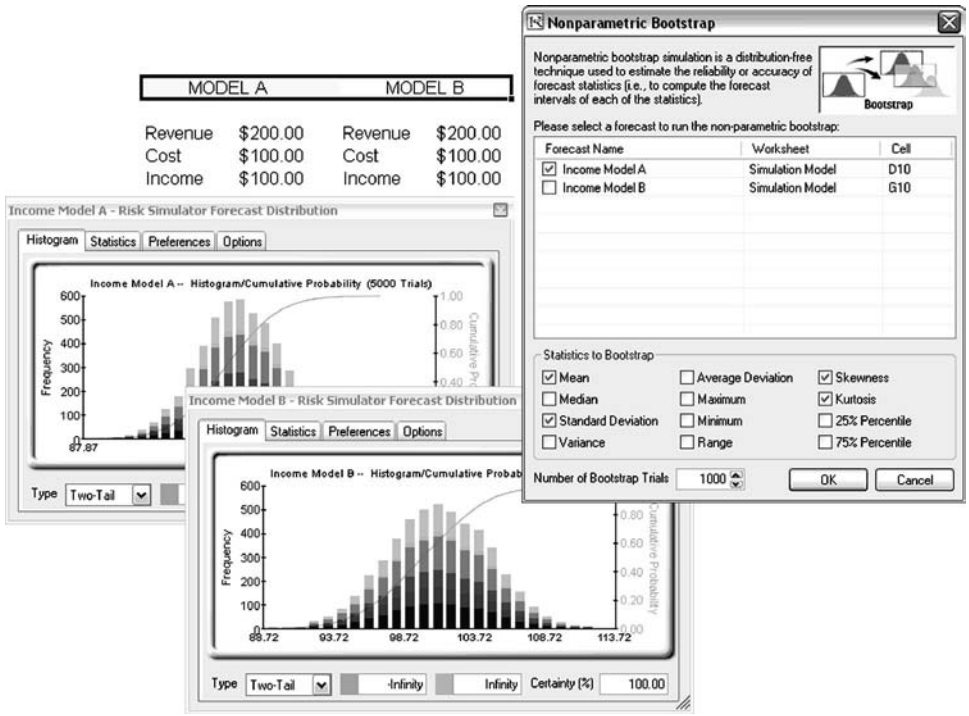


FIGURE I.34 Nonparametric bootstrap simulation

is not normally distributed or easily found, these classical methods are difficult to use. In contrast, bootstrapping analyzes sample statistics empirically by sampling the data repeatedly and creating distributions of the different statistics from each sampling. The classical methods of hypothesis testing are available in Risk Simulator and are explained in the next section. Classical methods provide higher power in their tests but rely on normality assumptions and can be used only to test the mean and variance of a distribution, as compared to bootstrap simulation, which provides lower power but is nonparametric and distribution-free, and can be used to test any distributional statistic.

PROCEDURE

1. Run a simulation with assumptions and forecasts.
2. Select **Risk Simulator | Tools | Nonparametric Bootstrap.**
3. Select only *one* forecast to bootstrap, select the statistic(s) to bootstrap, and enter the number of bootstrap trials and click **OK** (Figure I.34).

Results Interpretation

Figure I.35 illustrates some sample bootstrap results. The example file used was *Hypothesis Testing and Bootstrap Simulation*. For instance, the 90% confidence for the skewness statistic is between –0.0189 and 0.0952, such that the value 0 falls

within this confidence, indicating that on a 90% confidence, the skewness of this forecast is not statistically significantly different from zero, or that this distribution can be considered as symmetrical and not skewed. Conversely, if the value 0 falls outside of this confidence, then the opposite is true: The distribution is skewed (positively skewed if the forecast statistic is positive, and negatively skewed if the forecast statistic is negative).

Notes

The term *bootstrap* comes from the saying "to pull oneself up by one's own bootstraps" and is applicable because this method uses the distribution of statistics

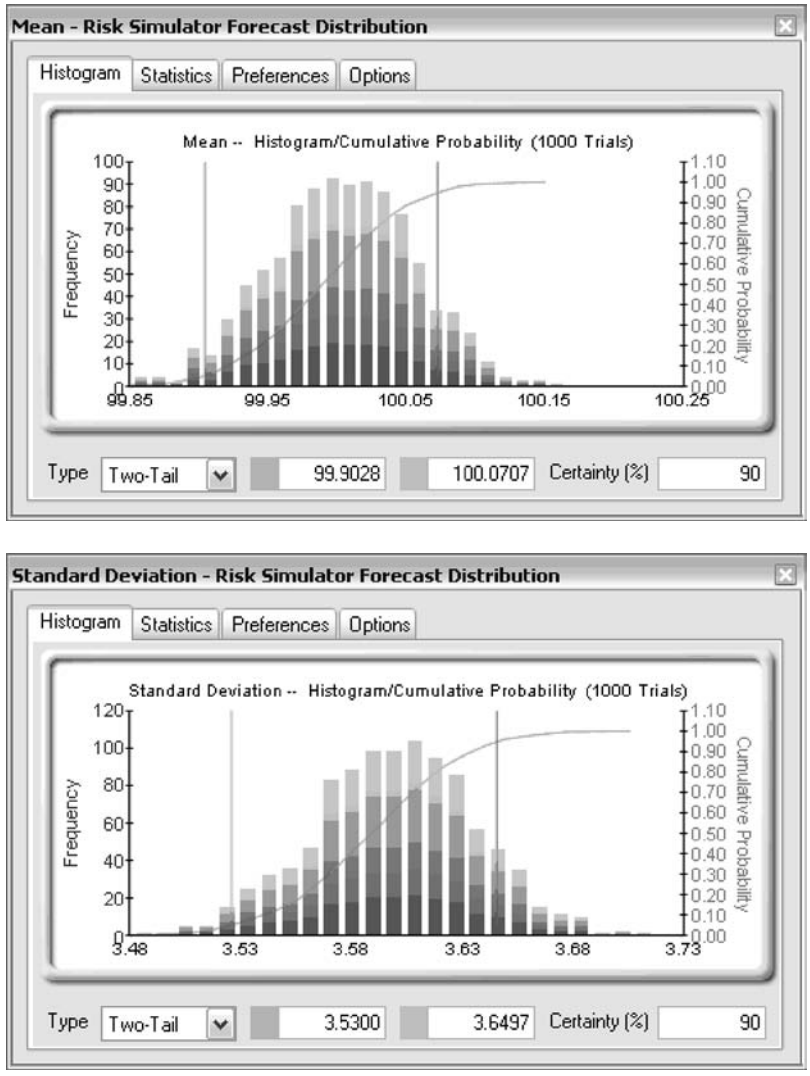


FIGURE I.35 Bootstrap simulation results (*Continued*)

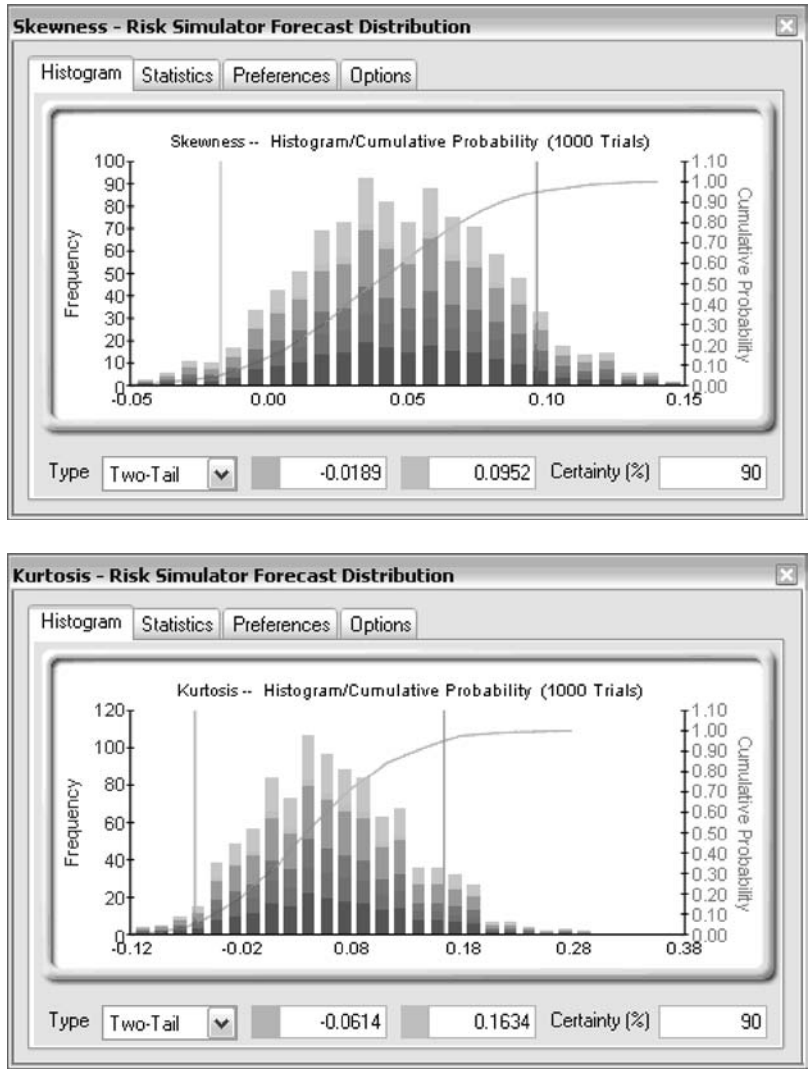


FIGURE I.35 (*Continued*)

themselves to analyze the accuracy of the statistics. Nonparametric simulation is simply randomly picking golf balls from a large basket with replacement where each golf ball is based on a historical data point. Suppose there are 365 golf balls in the basket (representing 365 historical data points). Imagine if you will that the value of each golf ball picked at random is written on a large whiteboard. The results of the 365 balls picked with replacement are written in the first column of the board with 365 rows of numbers. Relevant statistics (e.g., mean, median, mode, standard deviation, etc.) are calculated on these 365 rows. The process is then repeated, say, 5,000 times. The whiteboard will now be filled with 365 rows and 5,000 columns. Hence, 5,000 sets of statistics (i.e., there will be 5,000 means, 5,000 medians, 5,000 modes, 5,000 standard deviations, etc.) are tabulated and their distributions shown.

The relevant *statistics of the statistics* are then tabulated, where from these results you can ascertain how confident the simulated statistics are. Finally, bootstrap results are important because according to the Law of Large Numbers and Central Limit Theorem in statistics, the mean of the sample means is an unbiased estimator and approaches the true population mean when the sample size increases.

HYPOTHESIS TESTING

A hypothesis test is performed when testing the means and variances of two distributions to determine if they are statistically identical or statistically different from one another (i.e., to see if the differences between the means and variances of two different forecasts that occur are based on random chance or if they are, in fact, statistically significantly different from one another).

This analysis is related to bootstrap simulation with several differences. Classical hypothesis testing uses mathematical models and is based on theoretical distributions. This means that the precision and power of the test is higher than bootstrap simulation's empirically based method of simulating a simulation and letting the data tell the story. However, the classical hypothesis test is applicable only for testing means and variances of two distributions (and by extension, standard deviations) to see if they are statistically identical or different. In contrast, nonparametric bootstrap simulation can be used to test for any distributional statistics, making it more useful; the drawback is its lower testing power. Risk Simulator provides both techniques from which to choose.

PROCEDURE

1. Run a simulation.
2. Select **Risk Simulator | Tools | Hypothesis Testing.**
3. Select the two forecasts to test, select the type of hypothesis test you wish to run, and click **OK** (Figure I.36).

Results Interpretation

A two-tailed hypothesis test is performed on the null hypothesis (H_o) such that the population means of the two variables are statistically identical to one another. The alternative hypothesis (H_a) is such that the population means are statistically different from one another. If the calculated *p*-values are less than or equal to 0.01, 0.05, or 0.10 alpha test levels, it means that the null hypothesis is rejected, which implies that the forecast means are statistically significantly different at the 1%, 5%, and 10% significance levels. If the null hypothesis is not rejected when the *p*-values are high, the means of the two forecast distributions are statistically similar to one another. The same analysis is performed on variances of two forecasts at a time using the pairwise F-test. If the *p*-values are small, then the variances (and standard deviations) are statistically different from one another; otherwise, for large *p*-values, the variances are statistically identical to one another. The example file used was *Hypothesis Testing and Bootstrap Simulation.*

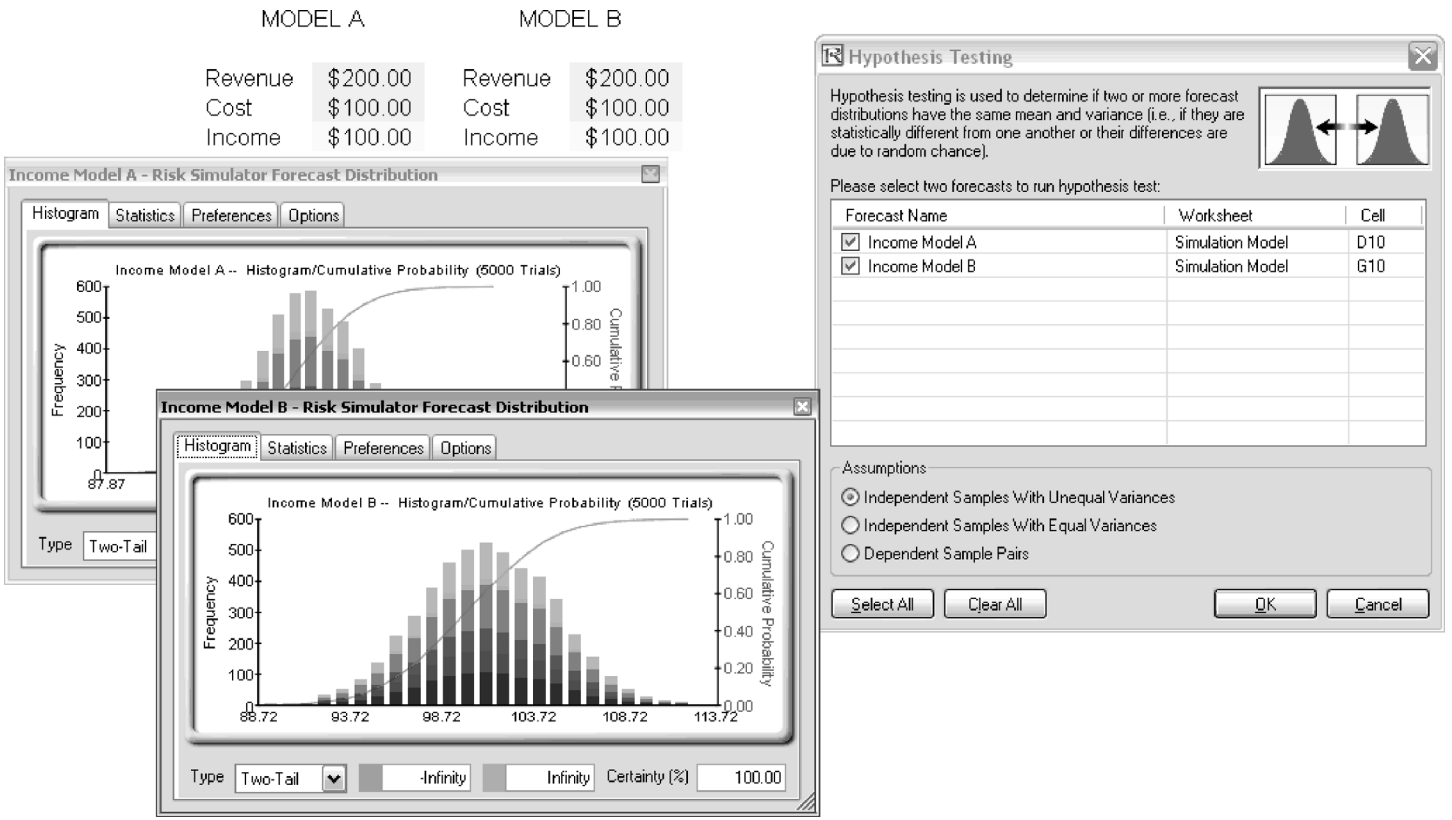


FIGURE 1.36 Hypothesis testing

Notes

The two-variable t-test with unequal variances (the population variance of forecast 1 is expected to be different from the population variance of forecast 2) is appropriate when the forecast distributions are from different populations (e.g., data collected from two different geographical locations, or two different operating business units, etc.). The two-variable t-test with equal variances (the population variance of forecast 1 is expected to be equal to the population variance of forecast 2) is appropriate when the forecast distributions are from similar populations (e.g., data collected from two different engine designs with similar specifications, etc.). The paired dependent two-variable t-test is appropriate when the forecast distributions are from exactly the same population and subjects (e.g., data collected from the same group of patients before an experimental drug was used and after the drug was applied, etc.).

DATA EXTRACTION, SAVING SIMULATION RESULTS, AND GENERATING REPORTS

Raw data of a simulation can be extracted very easily using Risk Simulator's *Data Extraction* routine. Both assumptions and forecasts can be extracted, but a simulation must be run first. The extracted data can then be used for a variety of other analyses, and the data can be extracted to different formats—for use in spreadsheets, databases, and other software products.

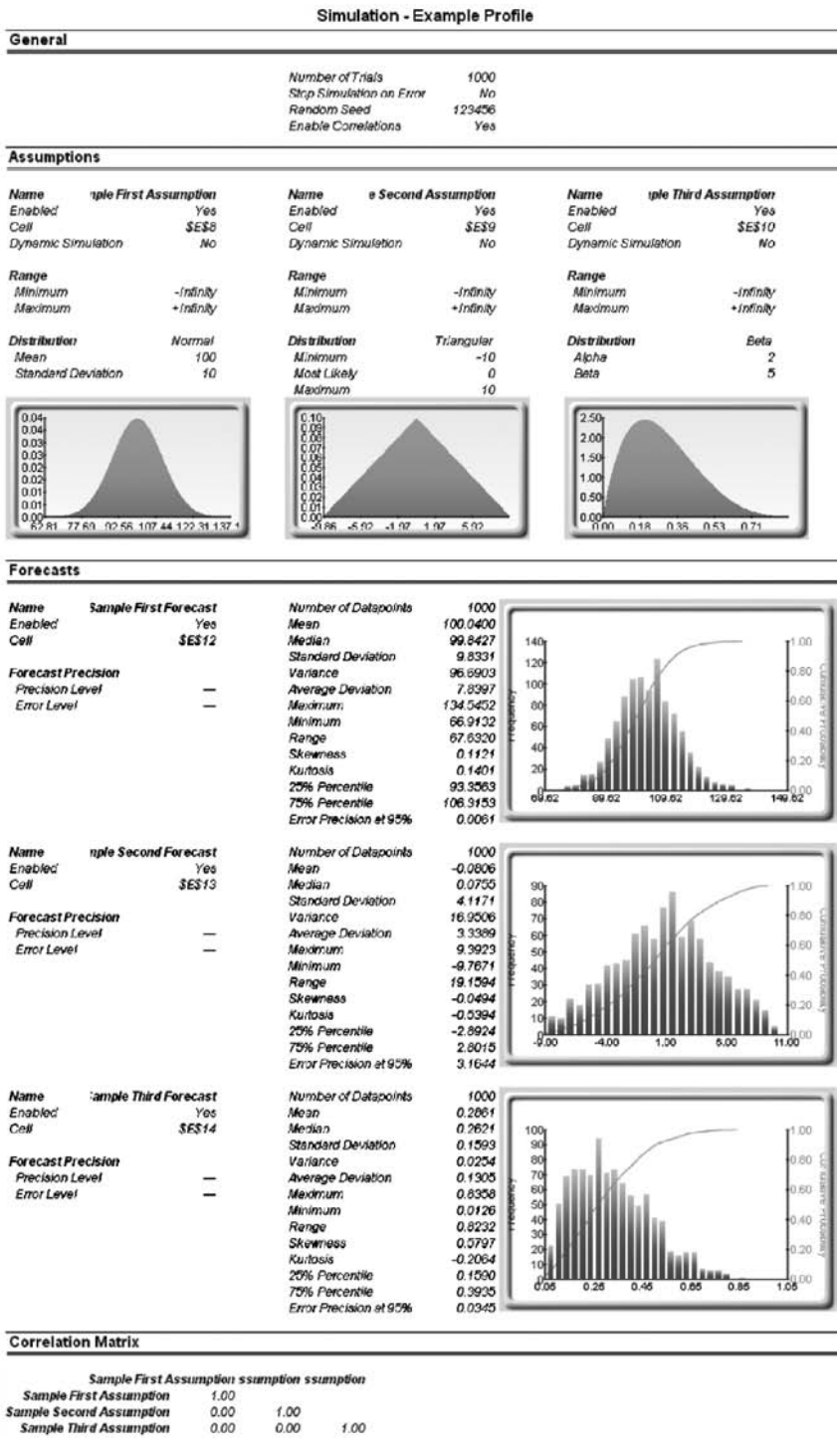
PROCEDURE

1. Open or create a model, define assumptions and forecasts, and run the simulation.
2. Select **Risk Simulator | Tools | Data Extraction.**
3. Select the assumptions and/or forecasts you wish to extract the data from and click **OK.**

The simulated data can be extracted to an Excel worksheet, a flat text file (for easy import into other software applications), or as *.risksim files (which can be reopened as Risk Simulator forecast charts at a later date). Finally, you can create a simulation report of all the assumptions and forecasts in the model by going to **Risk Simulator | Create Report.** A sample report is shown in Figure I.37.

REGRESSION AND FORECASTING DIAGNOSTIC TOOL

This advanced analytical tool in Risk Simulator is used to determine the econometric properties of your data. The diagnostics include checking the data for heteroskedasticity, nonlinearity, outliers, specification errors, micronumerosity, stationarity and stochastic properties, normality and sphericity of the errors, and multicollinearity. Each test is described in more detail in their respective reports in the model.



Simulation - Example Profile

General

Number of Trials	1000
Stop Simulation on Error	No
Random Seed	123456
Enable Correlations	Yes

Assumptions

Name	ple First Assumption
Enabled	Yes
Cell	E8
Dynamic Simulation	No
Range	
Minimum	-Infinity
Maximum	+Infinity
Distribution	Normal
Mean	100
Standard Deviation	10

Name	e Second Assumption
Enabled	Yes
Cell	E9
Dynamic Simulation	No
Range	
Minimum	-Infinity
Maximum	+Infinity
Distribution	Triangular
Minimum	-10
Most Likely	0
Maximum	10

Name	ple Third Assumption
Enabled	Yes
Cell	E10
Dynamic Simulation	No
Range	
Minimum	-Infinity
Maximum	+Infinity
Distribution	Beta
Alpha	2
Beta	5

Forecasts

Name	Sample First Forecast
Enabled	Yes
Cell	E12
Forecast Precision	
Precision Level	—
Error Level	—

Number of Datapoints	1000
Mean	100.0400
Median	99.8427
Standard Deviation	9.8331
Variance	96.6903
Average Deviation	7.8397
Maximum	134.5452
Minimum	66.9132
Range	67.6320
Skewness	0.1121
Kurtosis	0.1401
25% Percentile	93.3563
75% Percentile	106.9159
Error Precision at 95%	0.0061

Name	mple Second Forecast
Enabled	Yes
Cell	E13
Forecast Precision	
Precision Level	—
Error Level	—

Number of Datapoints	1000
Mean	-0.0806
Median	0.0755
Standard Deviation	4.1171
Variance	16.9506
Average Deviation	3.3389
Maximum	9.3923
Minimum	-9.7671
Range	19.1594
Skewness	-0.0494
Kurtosis	-0.5394
25% Percentile	-2.8924
75% Percentile	2.8015
Error Precision at 95%	3.1644

Name	ample Third Forecast
Enabled	Yes
Cell	E14
Forecast Precision	
Precision Level	—
Error Level	—

Number of Datapoints	1000
Mean	0.2861
Median	0.2621
Standard Deviation	0.1593
Variance	0.0254
Average Deviation	0.1305
Maximum	0.8358
Minimum	0.0126
Range	0.8232
Skewness	0.7797
Kurtosis	-0.2064
25% Percentile	0.1590
75% Percentile	0.3935
Error Precision at 95%	0.0345

Correlation Matrix

	Sample First Assumption	ssumption	ssumption
Sample First Assumption	1.00		
Sample Second Assumption	0.00	1.00	
Sample Third Assumption	0.00	0.00	1.00

FIGURE I.37 Sample simulation report

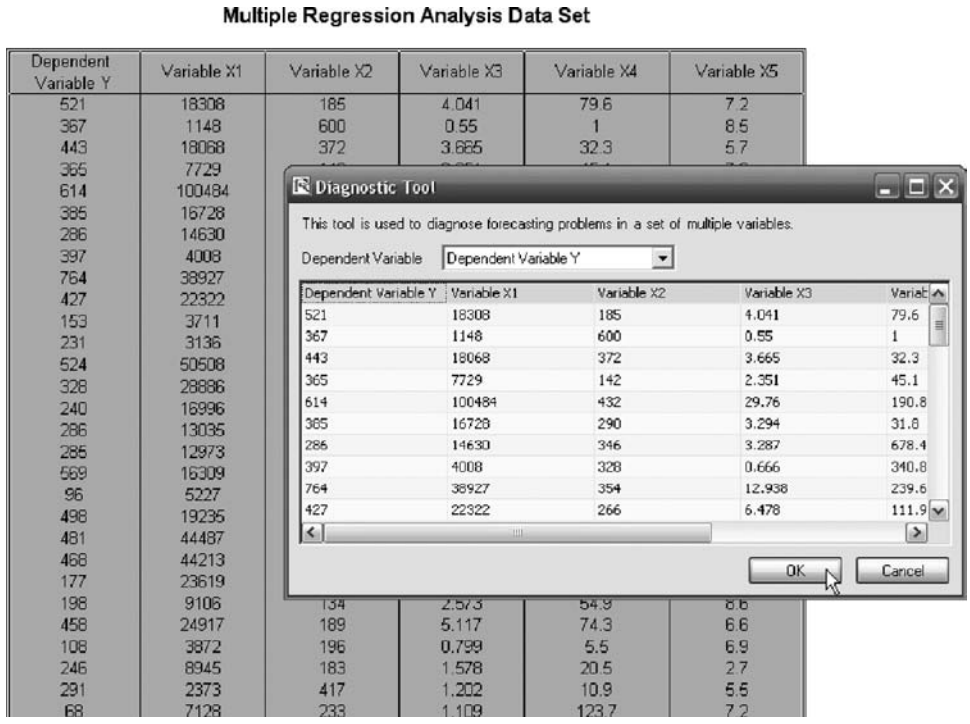


FIGURE I.38 Running the data diagnostic tool

PROCEDURE

- Open the example model (**Risk Simulator | Examples | Regression Diagnostics**) and go to the *Time-Series Data* worksheet and select the data including the variable names (cells **C5:H55**).
- Click on **Risk Simulator | Tools | Diagnostic Tool**.
- Check the data and select the *Dependent Variable Y* from the drop-down menu. Click **OK** when finished (Figure I.38).

A common violation in forecasting and regression analysis is heteroskedasticity; that is, the variance of the errors increases over time (see Figure I.39 for test results using the diagnostic tool). Visually, the width of the vertical data fluctuations increases or fans out over time, and typically, the coefficient of determination (R-squared coefficient) drops significantly when heteroskedasticity exists. If the variance of the dependent variable is not constant, then the error's variance will not be constant. Unless the heteroskedasticity of the dependent variable is pronounced, its effect will not be severe: the least-squares estimates will still be unbiased, and the estimates of the slope and intercept will either be normally distributed if the errors are normally distributed, or at least normally distributed asymptotically (as the number of data points becomes large) if the errors are not normally distributed. The estimate for the variance of the slope and overall variance will be inaccurate, but the inaccuracy is not likely to be substantial if the independent-variable values are symmetric about their mean.

Diagnostic Results

	Heteroskedasticity		Micronumerosity		Outliers		Nonlinearity	
Variable	W-Test p-value	Hypothesis Test result	Approximation result	Natural Lower Bound	Natural Upper Bound	Number of Potential Outliers	Nonlinear Test p-value	Hypothesis Test result
Y			no problems	-7.86	671.70	**2**		
Variable X1	0.2543	Homoskedastic	no problems	-21377.95	64713.03	**3**	0.2458	linear
Variable X2	0.3371	Homoskedastic	no problems	77.47	445.93	**2**	**0.0335**	**nonlinear**
Variable X3	0.3649	Homoskedastic	no problems	-5.77	15.69	**3**	**0.0305**	**nonlinear**
Variable X4	0.3066	Homoskedastic	no problems	-295.96	628.21	**4**	0.9298	linear
Variable X5	0.2495	Homoskedastic	no problems	3.35	9.38	**3**	0.2727	linear

FIGURE I.39 Results from tests of outliers, heteroskedasticity, micronumerosity, and nonlinearity

If the number of data points is small (micronumerosity), it may be difficult to detect assumption violations. With small sample sizes, assumption violations such as non-normality or heteroskedasticity of variances are difficult to detect even when they are present. With a small number of data points, linear regression offers less protection against violation of assumptions. With few data points, it may be hard to determine how well the fitted line matches the data, or whether a nonlinear function would be more appropriate. Even if none of the test assumptions are violated, a linear regression on a small number of data points may not have sufficient power to detect a significant difference between the slope and zero, even if the slope is nonzero. The power depends on the residual error, the observed variation in the independent variable, the selected significance alpha level of the test, and the number of data points. Power decreases as the residual variance increases, decreases as the significance level is decreased (i.e., as the test is made more stringent), increases as the variation in observed independent variable increases, and increases as the number of data points increases.

Values may not be identically distributed because of the presence of outliers. Outliers are anomalous values in the data. Outliers may have a strong influence over the fitted slope and intercept, giving a poor fit to the bulk of the data points. Outliers tend to increase the estimate of residual variance, lowering the chance of rejecting the null hypothesis (i.e., creating higher prediction errors). They may be due to recording errors, which may be correctable, or they may be due to the dependent-variable values not all being sampled from the same population. Apparent outliers may also be due to the dependent-variable values being from the same, but non-normal, population. However, a point may be an unusual value in either an independent or a dependent variable without necessarily being an outlier in the scatter plot. In regression analysis, the fitted line can be highly sensitive to outliers. In other words, least squares regression is not resistant to outliers; thus, neither is the fitted-slope estimate. A point vertically removed from the other points can cause the fitted line to pass close to it, instead of following the general linear trend of the rest of the data, especially if the point is relatively far horizontally from the center of the data.

However, great care should be taken when deciding if the outliers should be removed. Although in most cases when outliers are removed the regression results look better, *a priori* justification must first exist. For instance, if one is regressing the performance of a particular firm's stock returns, outliers caused by downturns in the stock market should be included; these are not truly outliers as they are inevitabilities in the business cycle. Forgoing these outliers and using the regression equation to forecast one's retirement fund based on the firm's stocks will yield incorrect results at best. In contrast, suppose the outliers are caused by a single nonrecurring business condition (e.g., merger and acquisition) and such business structural changes are not forecast to recur; then these outliers should be removed and the data cleansed prior to running a regression analysis. The analysis here only identifies outliers and it is up to the user to determine if they should remain or be excluded.

Sometimes, a nonlinear relationship between the dependent and independent variables is more appropriate than a linear relationship. In such cases, running a linear regression will not be optimal. If the linear model is not the correct form, then the slope and intercept estimates and the fitted values from the linear regression will be biased, and the fitted slope and intercept estimates will not be meaningful. Over a restricted range of independent or dependent variables, nonlinear models may be

well approximated by linear models (this is in fact the basis of linear interpolation), but for accurate prediction a model appropriate to the data should be selected. A nonlinear transformation should first be applied to the data before running a regression. One simple approach is to take the natural logarithm of the independent variable (other approaches include taking the square root or raising the independent variable to the second or third power) and run a regression or forecast using the nonlinearly transformed data.

Another typical issue when forecasting time-series data is whether the independent-variable values are truly independent of each other or are dependent. Dependent variable values collected over a time series may be autocorrelated. For serially correlated dependent-variable values, the estimates of the slope and intercept will be unbiased, but the estimates of their forecast and variances will not be reliable, and hence the validity of certain statistical goodness-of-fit tests will be flawed. For instance, interest rates, inflation rates, sales, revenues, and many other time-series data are typically autocorrelated, where the value in the current period is related to the value in a previous period, and so forth (clearly, the inflation rate in March is related to February's level, which in turn is related to January's level, and so forth). Ignoring such blatant relationships will yield biased and less accurate forecasts. In such events, an autocorrelated regression model or an ARIMA model may be better suited (**Risk Simulator | Forecasting | ARIMA**). Finally, the autocorrelation functions of a series that is nonstationary tend to decay slowly (see Nonstationary report in the model).

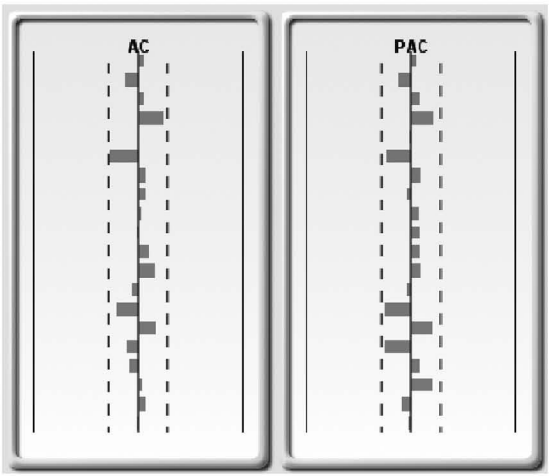
If autocorrelation AC(*1*) is nonzero, it means that the series is first-order serially correlated. If AC(k) dies off more or less geometrically with increasing lag, it implies that the series follows a low-order autoregressive process. If AC(k) drops to zero after a small number of lags, it implies that the series follows a low-order moving-average process. Partial correlation PAC(k) measures the correlation of values that are k periods apart after removing the correlation from the intervening lags. If the pattern of autocorrelation can be captured by an autoregression of order less than k, then the partial autocorrelation at lag k will be close to zero. Ljung-Box Q-statistics and their p-values at lag k have the null hypothesis that there is no autocorrelation up to order k. The dotted lines in the plots of the autocorrelations are the approximate two standard error bounds. If the autocorrelation is within these bounds, it is not significantly different from zero at the 5% significance level.

Autocorrelation measures the relationship to the past of the dependent Y variable to itself. Distributive lags, in contrast, are time-lag relationships between the dependent Y variable and different independent X variables. For instance, the movement and direction of mortgage rates tend to follow the federal funds rate but at a time lag (typically one to three months). Sometimes, time lags follow cycles and seasonality (e.g., ice cream sales tend to peak during the summer months and are hence related to the previous summer's sales, 12 months in the past). The distributive lag analysis (Figure I.40) shows how the dependent variable is related to each of the independent variables at various time lags, when all lags are considered simultaneously, to determine which time lags are statistically significant and should be considered.

Another requirement in running a regression model is the assumption of normality and sphericity of the error term. If the assumption of normality is violated or outliers are present, then the linear regression goodness-of-fit test may not be the most powerful or informative test available, and this could mean the difference between either detecting a linear fit or not. If the errors are not independent and not

Autocorrelation

Time Lag	AC	PAC	Lower Bound	Upper Bound	Q-Stat	Prob
1	0.0580	0.0580	-0.2828	0.2828	0.1786	0.6726
2	-0.1213	-0.1251	-0.2828	0.2828	0.9754	0.6140
3	0.0590	0.0756	-0.2828	0.2828	1.1679	0.7607
4	0.2423	0.2232	-0.2828	0.2828	4.4865	0.3442
5	0.0067	-0.0078	-0.2828	0.2828	4.4890	0.4814
6	-0.2654	-0.2345	-0.2828	0.2828	8.6516	0.1941
7	0.0814	0.0939	-0.2828	0.2828	9.0524	0.2489
8	0.0634	-0.0442	-0.2828	0.2828	9.3012	0.3175
9	0.0204	0.0873	-0.2828	0.2828	9.3278	0.4076
10	-0.0190	0.0865	-0.2828	0.2828	9.3512	0.4991
11	0.1035	0.0790	-0.2828	0.2828	10.0648	0.5246
12	0.1658	0.0978	-0.2828	0.2828	11.9466	0.4500
13	-0.0524	-0.0430	-0.2828	0.2828	12.1394	0.5162
14	-0.2050	-0.2523	-0.2828	0.2828	15.1738	0.3664
15	0.1782	0.2089	-0.2828	0.2828	17.5315	0.2881
16	-0.1022	-0.2591	-0.2828	0.2828	18.3296	0.3050
17	-0.0861	0.0808	-0.2828	0.2828	18.9141	0.3335
18	0.0418	0.1987	-0.2828	0.2828	19.0559	0.3884
19	0.0869	-0.0821	-0.2828	0.2828	19.6894	0.4135
20	-0.0091	-0.0269	-0.2828	0.2828	19.6966	0.4770



Distributive Lags

P-Values of Distributive Lag Periods of Each Independent Variable

Variable	1	2	3	4	5	6	7	8	9	10	11	12
X1	0.8467	0.2045	0.3336	0.9105	0.9757	0.1020	0.9205	0.1267	0.5431	0.9110	0.7495	0.4016
X2	0.6077	0.9900	0.8422	0.2851	**0.0638**	**0.0032**	0.8007	0.1551	0.4823	0.1126	**0.0519**	0.4383
X3	0.7394	0.2396	0.2741	0.8372	0.9808	**0.0464**	0.8355	**0.0545**	0.6828	0.7354	0.5093	0.3500
X4	**0.0061**	0.6739	0.7932	0.7719	0.6748	0.8627	0.5586	0.9046	0.5726	0.6304	0.4812	0.5707
X5	0.1591	0.2032	0.4123	0.5599	0.6418	0.3447	0.9190	0.9740	0.5185	0.2856	0.1489	0.7794

FIGURE I.40 Autocorrelation and distributive lag results

Test Result

Regression Error Average	0.00
Standard Deviation of Errors	141.83
D Statistic	0.1036
D Critical at 1%	0.1138
D Critical at 5%	0.1225
D Critical at 10%	0.1458

Null Hypothesis: The errors are normally distributed.

Conclusion: The errors are normally distributed at the 1% alpha level.

Errors	Relative Frequency	Observed	Expected	O-E
-219.04	0.02	0.02	0.0612	-0.0412
-202.53	0.02	0.04	0.0766	-0.0366
-186.04	0.02	0.06	0.0948	-0.0348
-174.17	0.02	0.08	0.1097	-0.0297
-162.13	0.02	0.10	0.1265	-0.0265
-161.62	0.02	0.12	0.1272	-0.0072
-160.39	0.02	0.14	0.1291	0.0109
-145.40	0.02	0.16	0.1526	0.0074
-138.92	0.02	0.18	0.1637	0.0163
-133.81	0.02	0.20	0.1727	0.0273
-120.76	0.02	0.22	0.1973	0.0227
-120.12	0.02	0.24	0.1985	0.0415

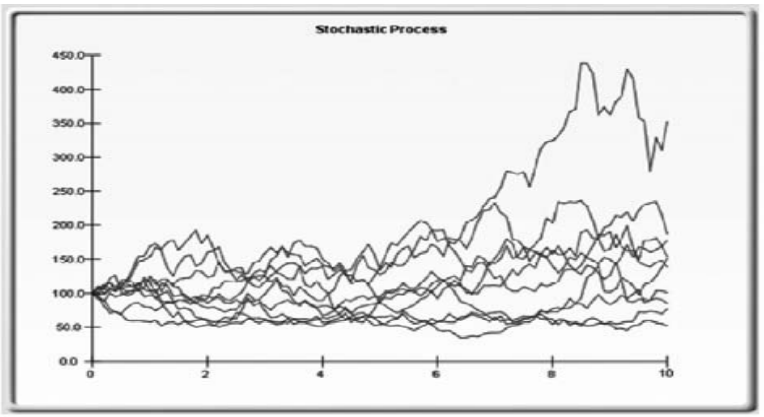
FIGURE I.41 Test for normality of errors

normally distributed, it may indicate that the data might be autocorrelated or suffer from nonlinearities or other more destructive errors. Independence of the errors can also be detected in the heteroskedasticity tests (Figure I.41).

The normality test on the errors performed is a nonparametric test, which makes no assumptions about the specific shape of the population from which the sample is drawn, allowing for smaller sample data sets to be analyzed. This test evaluates the null hypothesis of whether the sample errors were drawn from a normally distributed population, versus an alternate hypothesis that the data sample is not normally distributed. If the calculated D-Statistic is greater than or equal to the D-Critical values at various significance values, then reject the null hypothesis and accept the alternate hypothesis (the errors are not normally distributed). Otherwise, if the D-Statistic is less than the D-Critical value, do not reject the null hypothesis (the errors are normally distributed). This test relies on two cumulative frequencies: one derived from the sample data set, and the second from a theoretical distribution based on the mean and standard deviation of the sample data.

Sometimes, certain types of time-series data cannot be modeled using any other methods except for a stochastic process, because the underlying events are stochastic in nature. For instance, you cannot adequately model and forecast stock prices, interest rates, the price of oil, and other commodity prices using a simple regression model, because these variables are highly uncertain and volatile, and they do not follow a predefined static rule of behavior; in other words, the process is not stationary. Stationarity is checked here using the runs test, while another visual clue is found in the autocorrelation report (the ACF tends to decay slowly). A stochastic process is a sequence of events or paths generated by probabilistic laws. That is, random events can occur over time but are governed by specific statistical and probabilistic rules. The main stochastic processes include random walk or Brownian motion, mean-reversion, and jump-diffusion. These processes can be used to forecast a multitude of variables that seemingly follow random trends but are restricted by probabilistic laws. The process-generating equation is known in advance, but the actual results generated are unknown (Figure I.42).

The random walk or Brownian motion process can be used to forecast stock prices, prices of commodities, and other stochastic time-series data given a drift or growth rate and volatility around the drift path. The mean-reversion process can be



Statistical Summary

The following are the estimated parameters for a stochastic process given the data provided. It is up to you to determine if the probability of fit (similar to a goodness-of-fit computation) is sufficient to warrant the use of a stochastic process forecast, and if so, whether it is a random walk, mean-reversion, or a jump-diffusion model, or combinations thereof. In choosing the right stochastic process model, you will have to rely on past experiences and *a priori* economic and financial expectations of what the underlying data set is best represented by. These parameters can be entered into a stochastic process forecast (**Simulation | Forecasting | Stochastic Processes**).

Periodic

Drift Rate	*-1.48%*	*Reversion Rate*	*283.89%*	*Jump Rate*	*20.41%*
Volatility	*88.84%*	*Long-Term Value*	*327.72*	*Jump Size*	*237.89*

Probability of stochastic model fit: 46.48%

A high fit means a stochastic model is better than conventional models.

Runs	*20*	*Standard Normal*	*-1.7321*
Positive	*25*	*P-Value (1-tail)*	*0.0416*
Negative	*25*	*P-Value (2-tail)*	*0.0033*
Expected Run	*26*		

A low p-value (below 0.10, 0.05, 0.01) means that the sequence is not random and hence suffers from stationarity problems, and an ARIMA model might be more appropriate. Conversely, higher p-values indicate randomness and stochastic process models might be appropriate.

FIGURE I.42 Stochastic process parameter estimation

used to reduce the fluctuations of the random walk process by allowing the path to target a long-term value, making it useful for forecasting time-series variables that have a long-term rate such as interest rates and inflation rates (these are long-term target rates by regulatory authorities or the market). The jump-diffusion process is useful for forecasting time-series data when the variable can occasionally exhibit random jumps, such as oil prices or the price of electricity (discrete exogenous event shocks can make prices jump up or down). These processes can also be mixed and matched as required.

Multicollinearity exists when there is a linear relationship between the independent variables. When this occurs, the regression equation cannot be estimated at all. In near collinearity situations, the estimated regression equation will be biased and provide inaccurate results. This situation is especially true when a step-wise regression approach is used, where the statistically significant independent variables will be thrown out of the regression mix earlier than expected, resulting in a regression equation that is neither efficient nor accurate. One quick test of the presence of multicollinearity in a multiple regression equation is that the R-squared value is relatively high while the t-statistics are relatively low.

Another quick test is to create a correlation matrix between the independent variables. A high cross-correlation indicates a potential for autocorrelation. The rule of thumb is that a correlation with an absolute value greater than 0.75 is indicative of severe multicollinearity. Another test for multicollinearity is the use of the Variance Inflation Factor (VIF), obtained by regressing each independent variable to all the other independent variables, obtaining the R-squared value and calculating the VIF. A VIF exceeding 2.0 can be considered as severe multicollinearity. A VIF exceeding 10.0 indicates destructive multicollinearity (Figure I.43).

The Correlation Matrix lists the Pearson's product moment correlations (commonly referred to as the Pearson's *R*) between variable pairs. The correlation coefficient ranges between –1.0 and +1.0 inclusive. The sign indicates the direction of association between the variables, while the coefficient indicates the magnitude or strength of association. The Pearson's *R* measures only a linear relationship, and is less effective in measuring nonlinear relationships.

Correlation Matrix

CORRELATION	X2	X3	X4	X5
X1	0.333	0.959	0.242	0.237
X2	1.000	0.349	0.319	0.120
X3		1.000	0.196	0.227
X4			1.000	0.290

Variance Inflation Factor

VIF	X2	X3	X4	X5
X1	1.12	12.46	1.06	1.06
X2	N/A	1.14	1.11	1.01
X3		N/A	1.04	1.05
X4			N/A	1.09

FIGURE I.43 Multicollinearity errors

To test whether the correlations are significant, a two-tailed hypothesis test is performed and the resulting *p*-values are listed. *P*-values less than 0.10, 0.05, and 0.01 are highlighted to indicate statistical significance. In other words, a *p*-value for a correlation pair that is less than a given significance value is statistically significantly different from zero, indicating that there is a significant linear relationship between the two variables.

The Pearson's product moment correlation coefficient (R) between two variables (x and y) is related to the covariance (cov) measure where $R_{x,y} = \frac{COV_{x,y}}{s_x s_y}$. The benefit of dividing the covariance by the product of the two variables' standard deviations (s) is that the resulting correlation coefficient is bounded between –1.0 and +1.0 inclusive. This makes the correlation a good relative measure to compare among different variables (particularly with different units and magnitude). The Spearman rank-based nonparametric correlation is also included in the analysis. The Spearman's R is related to the Pearson's R in that the data is first ranked and then correlated. The rank correlations provide a better estimate of the relationship between two variables when one or both of them is nonlinear.

It must be stressed that a significant correlation does not imply causation. Associations between variables in no way imply that the change of one variable causes another variable to change. When two variables are moving independently of each other but in a related path, they may be correlated but their relationship might be spurious (e.g., a correlation between sunspots and the stock market might be strong, but one can surmise that there is no causality and that this relationship is purely spurious).

STATISTICAL ANALYSIS TOOL

Another very powerful tool in Risk Simulator is the statistical analysis tool, which determines the statistical properties of the data. The diagnostics run include checking the data for various statistical properties, from basic descriptive statistics to testing for and calibrating the stochastic properties of the data.

PROCEDURE

Open the example model (**Risk Simulator | Example Models | Statistical Analysis**) and go to the *Data* worksheet and select the data including the variable names (cells **C5:E55**).

- Click on **Risk Simulator | Tools | Statistical Analysis** (Figure I.44).
- Check the *data type*, whether the data selected is from a single variable or multiple variables arranged in columns. In our example, we assume that the data areas selected are from multiple variables. Click **OK** when finished.
- Choose the *statistical tests* you wish to perform. The suggestion (and by default) is to choose all the tests. Click **OK** when finished (Figure I.45).

Spend some time going through the reports generated to get a better understanding of the statistical tests performed (sample reports are shown in Figures I.46 to I.49).

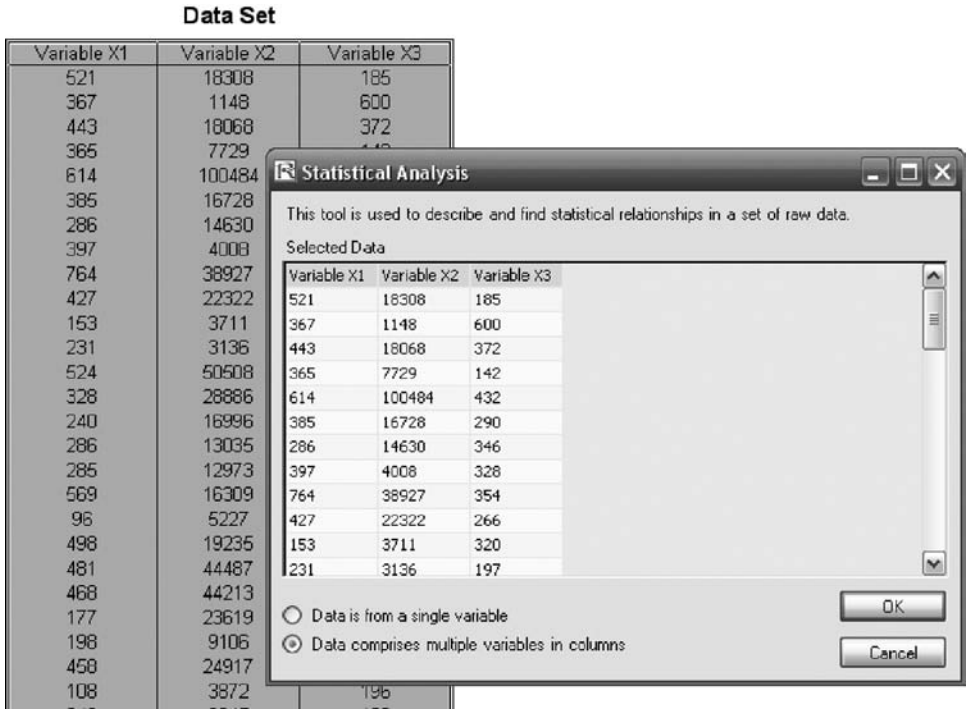


FIGURE I.44 Running the statistical analysis tool

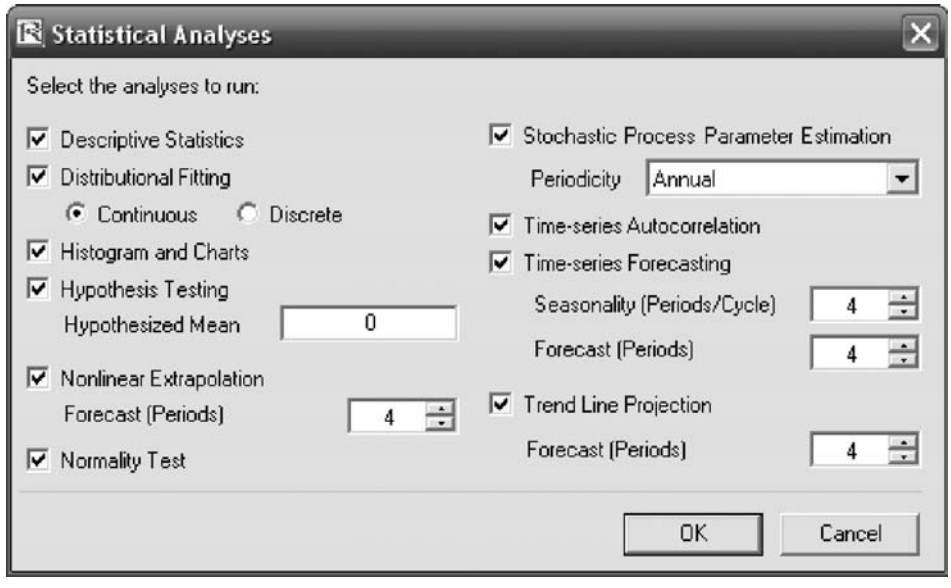


FIGURE I.45 Statistical tests

Descriptive Statistics

Analysis of Statistics

Almost all distributions can be described within 4 moments (some distributions require one moment, while others require two moments, and so forth). Descriptive statistics quantitatively capture these moments. The first moment describes the location of a distribution (i.e., mean, median, and mode) and is interpreted as the expected value, expected returns, or the average value of occurrences.

The Arithmetic Mean calculates the average of all occurrences by summing up all of the data points and dividing them by the number of points. The Geometric Mean is calculated by taking the power root of the products of all the data points and requires them to all be positive. The Geometric Mean is more accurate for percentages or rates that fluctuate significantly. For example, you can use Geometric Mean to calculate average growth rate given compound interest with variable rates. The Trimmed Mean calculates the arithmetic average of the data set after the extreme outliers have been trimmed. As averages are prone to significant bias when outliers exist, the Trimmed Mean reduces such bias in skewed distributions.

The Standard Error of the Mean calculates the error surrounding the sample mean. The larger the sample size, the smaller the error such that for an infinitely large sample size, the error approaches zero, indicating that the population parameter has been estimated. Due to sampling errors, the 95% Confidence Interval for the Mean is provided. Based on an analysis of the sample data points, the actual population mean should fall between these Lower and Upper Intervals for the Mean.

Median is the data point where 50% of all data points fall above this value and 50% below this value. Among the three first moment statistics, the median is least susceptible to outliers. A symmetrical distribution has the Median equal to the Arithmetic Mean. A skewed distribution exists when the Median is far away from the Mean. The Mode measures the most frequently occurring data point.

Minimum is the smallest value in the data set while Maximum is the largest value. Range is the difference between the Maximum and Minimum values.

The second moment measures a distribution's spread or width, and is frequently described using measures such as Standard Deviations, Variances, Quartiles, and Inter-Quartile Ranges. Standard Deviation indicates the average deviation of all data points from their mean. It is a popular measure as is associated with risk (higher standard deviations mean a wider distribution, higher risk, or wider dispersion of data points around the mean) and its units are identical to original data set's. The Sample Standard Deviation differs from the Population Standard Deviation in that the former uses a degree of freedom correction to account for small sample sizes. Also, Lower and Upper Confidence Intervals are provided for the Standard Deviation and the true population standard deviation falls within this interval. If your data set covers every element of the population, use the Population Standard Deviation instead. The two Variance measures are simply the squared values of the standard deviations.

The Coefficient of Variability is the standard deviation of the sample divided by the sample mean, proving a unit-free measure of dispersion that can be compared across different distributions (you can now compare distributions of values denominated in millions of dollars with one in billions of dollars, or meters and kilograms, etc.). The First Quartile measures the 25th percentile of the data points when arranged from its smallest to largest value. The Third Quartile is the value of the 75th percentile data point. Sometimes quartiles are used as the upper and lower ranges of a distribution as it truncates the data set to ignore outliers. The Inter-Quartile Range is the difference between the third and first quartiles, and is often used to measure the width of the center of a distribution.

Skewness is the third moment in a distribution. Skewness characterizes the degree of asymmetry of a distribution around its mean. Positive skewness indicates a distribution with an asymmetric tail extending toward more positive values. Negative skewness indicates a distribution with an asymmetric tail extending toward more negative values.

Kurtosis characterizes the relative peakedness or flatness of a distribution compared to the normal distribution. It is the fourth moment in a distribution. A positive Kurtosis value indicates a relatively peaked distribution. A negative kurtosis indicates a relatively flat distribution. The Kurtosis measured here has been centered to zero (certain other kurtosis measures are centered around 3.0). While both are equally valid, centering across zero makes the interpretation simpler. A high positive Kurtosis indicates a peaked distribution around its center and leptokurtic or fat tails. This indicates a higher probability of extreme events (e.g., catastrophic events, terrorist attacks, stock market crashes) than is predicted in a normal distribution.

Summary Statistics

Statistics	*Variable X1*		
Observations	*50.0000*	*Standard Deviation (Sample)*	*172.9140*
Arithmetic Mean	*331.9200*	*Standard Deviation (Population)*	*171.1761*
Geometric Mean	*281.3247*	*Lower Confidence Interval for Standard Deviation*	*148.6090*
Trimmed Mean	*325.1739*	*Upper Confidence Interval for Standard Deviation*	*207.7947*
Standard Error of Arithmetic Mean	*24.4537*	*Variance (Sample)*	*29899.2588*
Lower Confidence Interval for Mean	*283.0125*	*Variance (Population)*	*29301.2736*
Upper Confidence Interval for Mean	*380.8275*	*Coefficient of Variability*	*0.5210*
Median	*307.0000*	*First Quartile (Q1)*	*204.0000*
Mode	*47.0000*	*Third Quartile (Q3)*	*441.0000*
Minimum	*764.0000*	*Inter-Quartile Range*	*237.0000*
Maximum	*717.0000*	*Skewness*	*0.4838*
Range		*Kurtosis*	*-0.0952*

FIGURE I.46 Sample statistical analysis tool report

DISTRIBUTIONAL ANALYSIS TOOL

This is a statistical probability tool in Risk Simulator that is rather useful in a variety of settings, and can be used to compute the probability density function (PDF), which is also called the probability mass function (PMF) for discrete distributions (we will use these terms interchangeably), where given some distribution and its parameters, we can determine the probability of occurrence given some outcome x. In addition, the cumulative distribution function (CDF) can also be computed, which is the sum of the PDF values up to and including this x value. Finally, the inverse cumulative distribution function (ICDF) is used to compute the value x given the probability of occurrence.

This tool is accessible via **Risk Simulator | Tools | Distributional Analysis**. As an example, Figure I.50 shows the computation of a binomial distribution (i.e., a distribution with two outcomes, such as the tossing of a coin, where the outcome is either heads or tails, with some prescribed probability of heads and tails). Suppose we toss a coin two times and set the outcome heads as a success; we use the binomial distribution with Trials = 2 (tossing the coin twice) and Probability = 0.50 (the probability of success, of getting heads). Selecting the PDF and setting the range of

Hypothesis Test (t-Test on the Population Mean of One Variable)

Statistical Summary

Statistics from Dataset:		*Calculated Statistics:*	
Observations	50	*t-Statistic*	13.5734
Sample Mean	331.92	*P-Value (right-tail)*	**0.0000**
Sample Standard Deviation	172.91	*P-Value (left-tailed)*	**1.0000**
		P-Value (two tailed)	**0.0000**
User Provided Statistics:			
		Null Hypothesis (Ho):	*μ = Hypothesized Mean*
Hypothesized Mean	0.00	*Alternate Hypothesis (Ha):*	*μ <> Hypothesized Mean*

Notes: "<>" denotes "greater than" for right-tail, "less than" for left-tail, or "not equal to" for two-tail hypothesis tests.

Hypothesis Testing Summary

The one-variable t-test is appropriate when the population standard deviation is not known but the sampling distribution is assumed to be approximately normal (the t-test is used when the sample size is less than 30 but is also appropriate and in fact, provides more conservative results with larger data sets). This t-test can be applied to three types of hypothesis tests: a two-tailed test, a right-tailed test, and a left-tailed test. All three tests and their respective results are listed below for your reference.

Two-Tailed Hypothesis Test

A two-tailed hypothesis tests the null hypothesis Ho such that the population mean is statistically identical to the hypothesized mean. The alternative hypothesis is that the real population mean is statistically different from the hypothesized mean when tested using the sample dataset. Using a t-test, if the computed p-value is less than a specified significance amount (typically 0.10, 0.05, or 0.01), this means that the population mean is statistically significantly different than the hypothesized mean at **10%, 5% and 1%** significance value (or at the **90%, 95%, and 99%** statistical confidence). Conversely, if the p value is higher than 0.10, 0.05, or 0.01, the population mean is statistically identical to the hypothesized mean and any differences are due to random chance.

Right-Tailed Hypothesis Test

A right-tailed hypothesis tests the null hypothesis Ho such that the population mean is statistically less than or equal to the hypothesized mean. The alternative hypothesis is that the real population mean is statistically greater than the hypothesized mean when tested using the sample dataset. Using a t-test, if the p-value is less than a specified significance amount (typically 0.10, 0.05, or 0.01), this means that the population mean is statistically significantly greater than the hypothesized mean at **10%, 5% and 1%** significance value (or **90%, 95%, and 99%** statistical confidence). Conversely, if the p value is higher than 0.10, 0.05, or 0.01, the population mean is statistically similar or less than the hypothesized mean.

Left-Tailed Hypothesis Test

A left-tailed hypothesis tests the null hypothesis Ho such that the population mean is statistically greater than or equal to the hypothesized mean. The alternative hypothesis is that the real population mean is statistically less than the hypothesized mean when tested using the sample dataset. Using a t-test, if the p-value is less than a specified significance amount (typically 0.10, 0.05, or 0.01), this means that the population mean is statistically significantly less than the hypothesized mean at **10%, 5%, and 1%** significance value (or **90%, 95%, and 99%** statistical confidence). Conversely, if the p-value is higher than 0.10, 0.05, or 0.01, the population mean is statistically similar or greater than the hypothesized mean and any differences are due to random chance.

Because the t-test is more conservative and does not require a known population standard deviation as in the Z-test, we only use this t-test.

FIGURE I.47 Sample statistical analysis tool report (hypothesis testing of one variable)

values x as from 0 to 2 with a step size of 1 (this means we are requesting the values 0, 1, and 2 for x), the resulting probabilities are provided in the table and graphically, as well as the theoretical four moments of the distribution. As the outcomes of the coin toss are heads-heads, tails-tails, heads-tails, and tails-heads, the probability of getting exactly no heads is 25%, one heads is 50%, and two heads is 25%.

Similarly, we can obtain the exact probabilities of tossing the coin, say, 20 times, as seen in Figure I.51. The results are presented in both table and graphical formats.

Figure I.51 shows the same binomial distribution but now the CDF is computed. The CDF is simply the sum of the PDF values up to the point x. For instance, in Figure I.51, we see that the probabilities of 0, 1, and 2 are 0.000001, 0.000019, and 0.000181, whose sum is 0.000201, which is the value of the CDF at $x = 2$ in Figure I.52. Whereas the PDF computes the probabilities of getting two heads, the CDF computes the probability of getting no more than two heads (or probabilities of 0, 1, and 2 heads). Taking the complement (i.e., 1 – 0.000201 obtains 0.999799 or 99.9799%) provides the probability of getting three heads or more.

Using this distributional analysis tool, even more advanced distributions can be analyzed, such as the gamma, beta, negative binomial, and many others in Risk

Test for Normality

The Normality test is a form of nonparametric test, which makes no assumptions about the specific shape of the population from which the sample is drawn, allowing for smaller sample data sets to be analyzed. This test evaluates the null hypothesis of whether the data sample was drawn from a normally distributed population, versus an alternate hypothesis that the data sample is not normally distributed. If the calculated p-value is less than or equal to the alpha significance value then reject the null hypothesis and accept the alternate hypothesis. Otherwise, if the p-value is higher than the alpha significance value, do not reject the null hypothesis. This test relies on two cumulative frequencies: one derived from the sample data set, the second from a theoretical distribution based on the mean and standard deviation of the sample data. An alternative to this test is the Chi-Square test for normality. The Chi-Square test requires more data points to run compared to the Normality test used here.

Test Result

Data Average	331.92
Standard Deviation	172.91
D Statistic	0.0659
D Critical at 1%	0.1150
D Critical at 5%	0.1237
D Critical at 10%	0.1473

Null Hypothesis: The data is normally distributed.

Conclusion: The sample data is normally distributed at the 1% alpha level.

Data	Relative Frequency	Observed	Expected	O-E
47.00	0.02	0.02	0.0497	-0.0297
68.00	0.02	0.04	0.0635	-0.0235
87.00	0.02	0.06	0.0783	-0.0183
96.00	0.02	0.08	0.0862	-0.0062
102.00	0.02	0.10	0.0918	0.0082
108.00	0.02	0.12	0.0977	0.0223
114.00	0.02	0.14	0.1038	0.0362
127.00	0.02	0.16	0.1180	0.0420
153.00	0.02	0.18	0.1504	0.0296
177.00	0.02	0.20	0.1851	0.0149
186.00	0.02	0.22	0.1994	0.0206
188.00	0.02	0.24	0.2026	0.0374
198.00	0.02	0.26	0.2193	0.0407
222.00	0.02	0.28	0.2625	0.0175
231.00	0.02	0.30	0.2797	0.0203
240.00	0.02	0.32	0.2975	0.0225
246.00	0.02	0.34	0.3096	0.0304
251.00	0.02	0.36	0.3199	0.0401
265.00	0.02	0.38	0.3494	0.0306
280.00	0.02	0.40	0.3820	0.0180
285.00	0.02	0.42	0.3931	0.0269
286.00	0.04	0.46	0.3953	0.0647
291.00	0.02	0.48	0.4065	0.0735
303.00	0.02	0.50	0.4336	0.0664
311.00	0.02	0.52	0.4519	0.0681

FIGURE I.48 Sample statistical analysis tool report (normality test)

Simulator. As a further example of the tool's use in a continuous distribution and the ICDF functionality, Figure I.53 shows the standard normal distribution (normal distribution with a mean of zero and standard deviation of one), where we apply the ICDF to find the value of x that corresponds to the cumulative probability of 97.50% (CDF). That is, a one-tail CDF of 97.50% is equivalent to a two-tail 95% confidence interval (there is a 2.50% probability in the right tail and 2.50% in the left tail, leaving 95% in the center or confidence interval area, which is equivalent to a 97.50% area for one tail). The result is the familiar Z-Score of 1.96. Therefore, using this distributional analysis tool, and the standardized scores for other distributions, the exact and cumulative probabilities of other distributions can all be obtained quickly and easily.

PORTFOLIO OPTIMIZATION

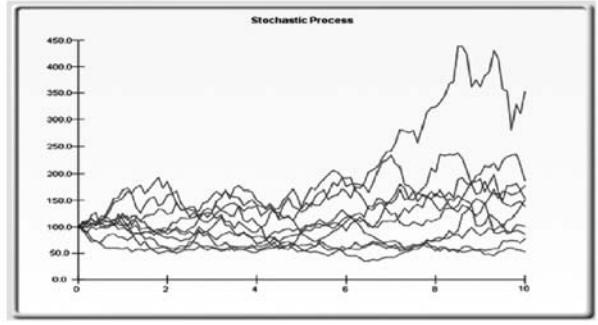
In today's competitive global economy, companies are faced with many difficult decisions. These decisions include allocating financial resources, building or expanding facilities, managing inventories, and determining product-mix strategies. Such

Stochastic Process - Parameter Estimations

Statistical Summary

A stochastic process is a sequence of events or paths generated by probabilistic laws. That is, random events can occur over time but are governed by specific statistical and probabilistic rules. The main stochastic processes include Random Walk or Brownian Motion, Mean-Reversion, and Jump-Diffusion. These processes can be used to forecast a multitude of variables that seemingly follow random trends but yet are restricted by probabilistic laws. The process-generating equation is known in advance but the actual results generated is unknown.

The Random Walk Brownian Motion process can be used to forecast stock prices, prices of commodities, and other stochastic time-series data given a drift or growth rate and a volatility around the drift path. The Mean-Reversion process can be used to reduce the fluctuations of the Random Walk process by allowing the path to target a long-term value, making it useful for forecasting time-series variables that have a long-term rate such as interest rates and inflation rates (these are long-term target rates by regulatory authorities or the market). The Jump-Diffusion process is useful for forecasting time-series data when the variable can occasionally exhibit random jumps, such as oil prices or price of electricity (discrete exogenous event shocks can make prices jump up or down). Finally, these three stochastic processes can be mixed and matched as required.



Statistical Summary

The following are the estimated parameters for a stochastic process given the data provided. It is up to you to determine if the probability of fit (similar to a goodness-of-fit computation) is sufficient to warrant the use of a stochastic process forecast, and if so, whether it is a random walk, mean-reversion, or a jump-diffusion model, or combinations thereof. In choosing the right stochastic process model, you will have to rely on past experiences and a priori economic and financial expectations of what the underlying data set is best represented by. These parameters can be entered into a stochastic process forecast (**Simulation** | **Forecasting** | **Stochastic Processes**).

(Annualized)

Drift Rate	*5.86%*	*Reversion Rate*	*N/A*	*Jump Rate*	*16.33%*
Volatility	*7.04%*	*Long-Term Value*	*N/A*	*Jump Size*	*21.33*

Probability of stochastic model fit: *4.63%*

FIGURE I.49 Sample statistical analysis tool report (stochastic parameter estimation)

decisions might involve thousands or millions of potential alternatives. Considering and evaluating each of them would be impractical or even impossible. A model can provide valuable assistance in incorporating relevant variables when analyzing decisions and finding the best solutions for making decisions. Models capture the most important features of a problem and present them in a form that is easy to interpret. Models often provide insights that intuition alone cannot. An optimization model has three major elements: decision variables, constraints, and an objective. In short, the optimization methodology finds the best combination or permutation of decision variables (e.g., which products to sell and which projects to execute) in every conceivable way such that the objective is maximized (e.g., revenues and net income) or minimized (e.g., risk and costs) while still satisfying the constraints (e.g., budget and resources).

Obtaining optimal values generally requires that you search in an iterative or ad hoc fashion. This search involves running one iteration for an initial set of values, analyzing the results, changing one or more values, rerunning the model, and repeating the process until you find a satisfactory solution. This process can be very tedious and time-consuming even for small models, and often it is not clear how to adjust the values from one iteration to the next.

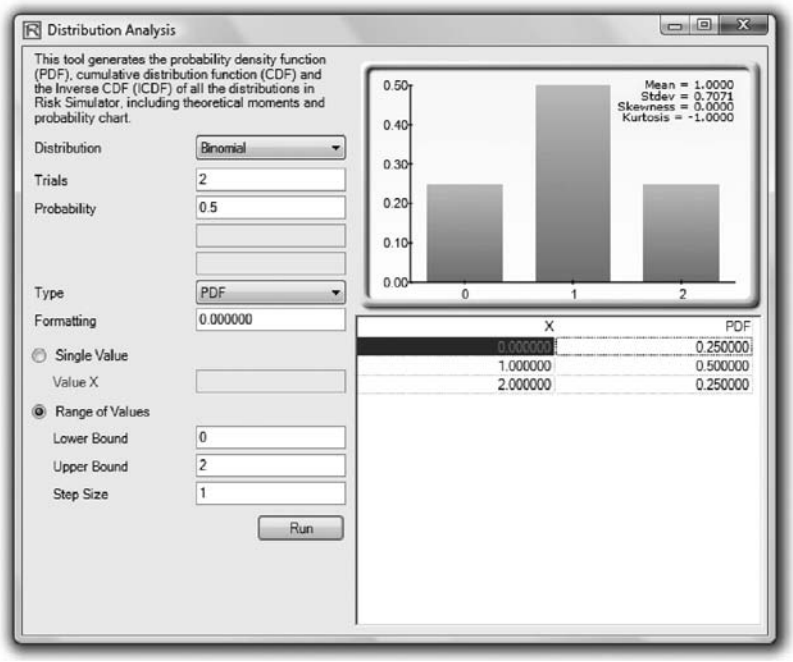


FIGURE I.50 Distributional analysis tool (binomial distribution with 2 trials and a 0.5 probability of success)

A more rigorous method systematically enumerates all possible alternatives. This approach guarantees optimal solutions if the model is correctly specified. Suppose that an optimization model depends on only two decision variables. If each variable has 10 possible values, trying each combination requires 100 iterations (10^2 alternatives). If each iteration is very short (e.g., 2 seconds), then the entire process could be done in approximately three minutes of computer time.

However, instead of two decision variables, consider six, then consider that trying all combinations requires 1,000,000 iterations (10^6 alternatives). It is easily possible for complete enumeration to take weeks, months, or even years to carry out.

The Traveling Financial Planner

A very simple example is in order. Table I.1 illustrates the traveling financial planner problem. Suppose the traveling financial planner has to make three sales trips to New York, Chicago, and Seattle. Further suppose that the order of arrival at each city is irrelevant. All that is important in this simple example is to find the lowest total cost possible to cover all three cities. Table I.1 also lists the flight costs from these different cities.

The problem here is cost minimization, suitable for optimization. One basic approach to solving this problem is through an ad hoc or brute force method. That is, manually list all six possible permutations, as seen in Table I.2. Clearly the cheapest itinerary is going from the East Coast to the West Coast, going from New York to

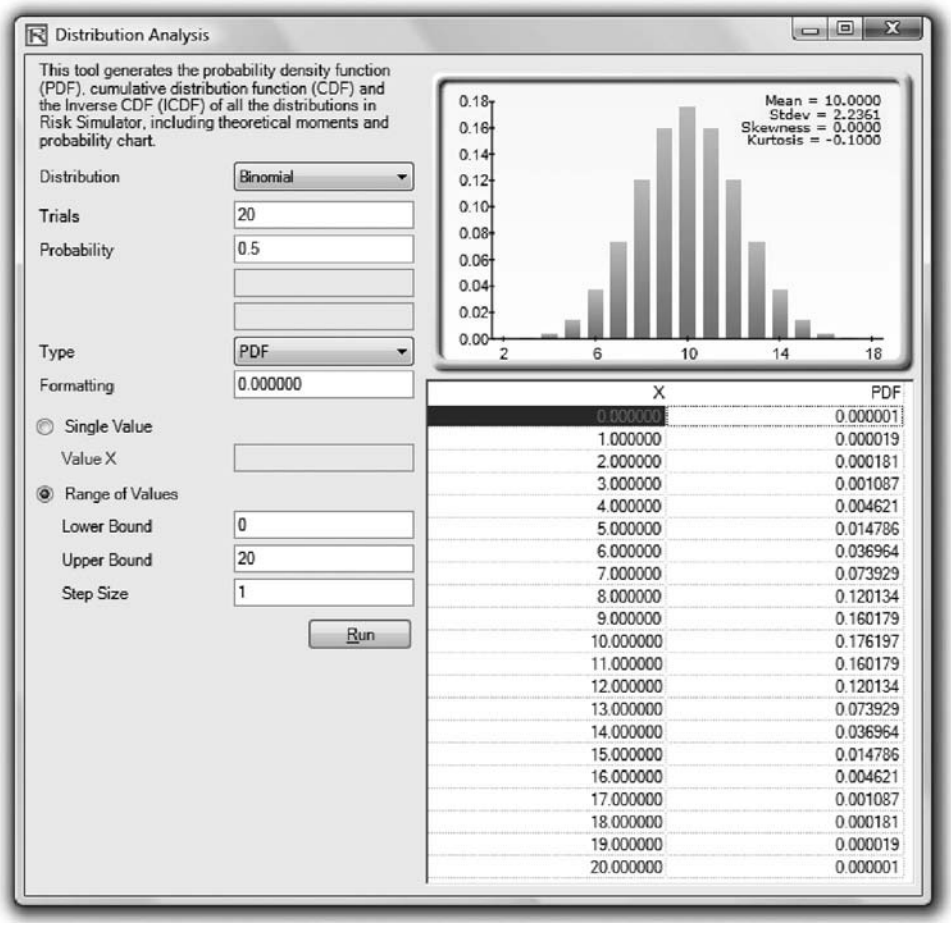


FIGURE I.51 Distributional analysis tool (binomial distribution with 20 trials)

Chicago and finally on to Seattle. Here, the problem is simple and can be calculated manually, as there were three cities and hence six possible itineraries. However, add two more cities and the total number of possible itineraries jumps to 120. Performing an ad hoc calculation will be fairly intimidating and time-consuming. On a larger scale, suppose there are 100 cities on the salesperson's list; the possible itineraries will be as many as 9.3×10^{157}. The problem will take many years to calculate manually, which is where optimization software steps in, automating the search for the optimal itinerary.

The example illustrated up to now is a deterministic optimization problem; that is, the airline ticket prices are known ahead of time and are assumed to be constant. Now suppose the ticket prices are not constant but are uncertain, following some distribution (e.g., a ticket from Chicago to Seattle averages $325, but is never cheaper than $300 and usually doesn't exceed $500). The same uncertainty applies to tickets for the other cities. The problem now becomes an *optimization under uncertainty*. Ad hoc and brute force approaches simply do not work. Software such as Risk

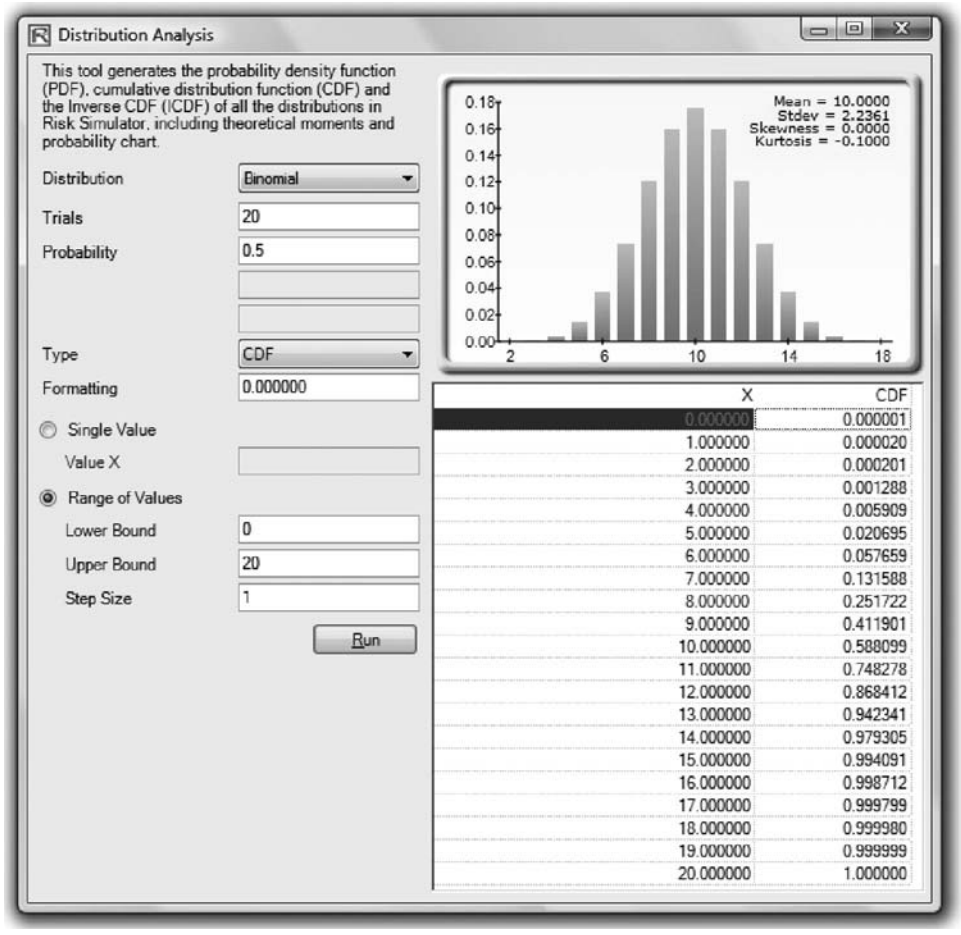


FIGURE 1.52 Distributional analysis tool (binomial distribution's CDF with 20 trials)

Simulator can take over this optimization problem and automate the entire process seamlessly. The next section discusses the terms required in an optimization under uncertainty.

The Lingo of Optimization

Before embarking on solving an optimization problem, it is vital to understand the terminology of optimization—the terms used to describe certain attributes of the optimization process. These words include *decision variables*, *constraints*, and *objectives*.

Decision variables are quantities over which you have control—for example, the amount of a product to make, the number of dollars to allocate among different investments, or which projects to select from among a limited set. As an example, portfolio optimization analysis includes a go or no-go decision on particular projects.

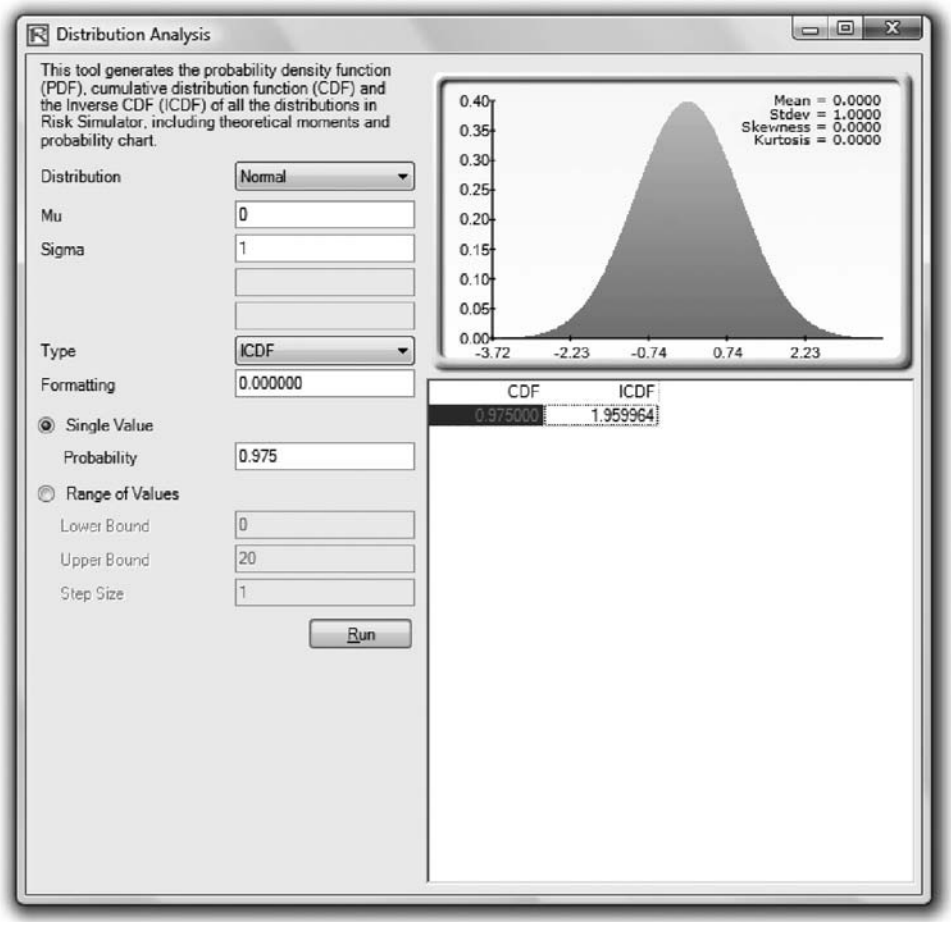


FIGURE I.53 Distributional Analysis Tool (Normal Distribution's ICDF and Z-Score)

In addition, the dollar or percentage budget allocation across multiple projects also can be structured as decision variables.

Constraints describe relationships among decision variables that restrict the values of the decision variables. For example, a constraint might ensure that the total amount of money allocated among various investments cannot exceed a specified

TABLE I.1 Traveling Financial Planner

Seattle to Chicago	$325
Chicago to Seattle	$225
New York to Seattle	$350
Seattle to New York	$375
Chicago to New York	$325
New York to Chicago	$325

TABLE I.2 Multiple Combinations of the Traveling Financial Planner Problem

Seattle–Chicago–New York	$\$325 + \$325 = \$650$
Seattle–New York–Chicago	$\$375 + \$325 = \$700$
Chicago–Seattle–New York	$\$225 + \$375 = \$600$
Chicago–New York–Seattle	$\$325 + \$350 = \$675$
New York–Seattle–Chicago	$\$350 + \$325 = \$675$
New York–Chicago–Seattle	$\$325 + \$225 = \$550$

Three cities means $3! = 3 \times 2 \times 1 = 6$ itinerary permutations.
Five cities means $5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$ permutations.
One hundred cities means $100! = 100 \times 99 \times \cdots \times 1 = 9.3 \times 10^{157}$ permutations!

amount (or at most one project from a certain group can be selected), budget constraints, timing restrictions, minimum returns, or risk tolerance levels.

Objectives give a mathematical representation of the model's desired outcome, such as maximizing profit or minimizing cost, in terms of the decision variables. In financial analysis, for example, the objective may be to maximize returns while minimizing risks (maximizing the Sharpe ratio or returns-to-risk ratio).

The solution to an optimization model provides a set of values for the decision variables that optimizes (maximizes or minimizes) the associated objective. If the real business conditions were simple and if the future were predictable, all data in an optimization model would be constant, making the model deterministic.

In many cases, however, a deterministic optimization model cannot capture all the relevant intricacies of a practical decision-making environment. When a model's data are uncertain and can only be described probabilistically, the objective will have some probability distribution for any chosen set of decision variables. You can find this probability distribution by simulating the model using Risk Simulator. An optimization model under uncertainty has several additional elements, including assumptions and forecasts.

Assumptions capture the uncertainty of model data using probability distributions, whereas forecasts are the frequency distributions of possible results for the model. Forecast statistics are summary values of a forecast distribution, such as the mean, standard deviation, and variance. The optimization process controls the optimization by maximizing or minimizing the objective.

Each optimization model has one objective, a variable that mathematically represents the model's objective in terms of the assumption and decision variables. Optimization's job is to find the optimal (minimum or maximum) value of the objective by selecting and improving different values for the decision variables. When model data are uncertain and can only be described using probability distributions, the objective itself will have some probability distribution for any set of decision variables.

Many algorithms exist to run optimization, and many different procedures exist when optimization is coupled with Monte Carlo simulation. In Risk Simulator, there are three distinct optimization procedures and optimization types as well as different decision variable types. For instance, Risk Simulator can handle ***continuous decision variables*** (1.2535, 0.2215, and so forth) as well as ***integers decision variables*** (e.g., 1, 2, 3, 4, and so forth), ***binary decision variables*** (1 and 0 for go and no-go decisions), and ***mixed decision variables*** (both integers and continuous variables).

On top of that, Risk Simulator can handle ***linear optimization*** (i.e., when both the objective and constraints are all linear equations and functions) as well as ***nonlinear optimizations*** (i.e., when the objective and constraints are a mixture of linear as well as nonlinear functions and equations).

As far as the optimization process is concerned, Risk Simulator can be used to run a ***static optimization***; that is, an optimization that is run on a static model, where no simulations are run. In other words, all the inputs in the model are static and unchanging. This optimization type is applicable when the model is assumed to be known and no uncertainties exist. Also, a static optimization can be first run to determine the optimal portfolio and its corresponding optimal allocation of decision variables before more advanced optimization procedures are applied. For instance, before running a stochastic optimization problem, a static optimization is first run to determine if there exist solutions to the optimization problem before a more protracted analysis is performed.

Next, ***dynamic optimization*** is applied when Monte Carlo simulation is used together with optimization. Another name for such a procedure is ***simulation-optimization***. That is, a simulation is first run; the results of the simulation are then applied in the Excel model, and then an optimization is applied to the simulated values. In other words, a simulation is run for N trials, and then an optimization process is run for M iterations until the optimal results are obtained or an infeasible set is found. That is, using Risk Simulator's optimization module, you can choose which forecast and assumption statistics to use and replace in the model after the simulation is run. Then, these forecast statistics can be applied in the optimization process. This approach is useful when you have a large model with many interacting assumptions and forecasts, and when some of the forecast statistics are required in the optimization. For example, if the standard deviation of an assumption or forecast is required in the optimization model (e.g., computing the Sharpe ratio in asset allocation and optimization problems where we have mean divided by standard deviation of the portfolio), then this approach should be used.

The ***stochastic optimization*** process, in contrast, is similar to the dynamic optimization procedure with the exception that the entire dynamic optimization process is repeated T times. That is, a simulation with N trials is run, and then an optimization is run with M iterations to obtain the optimal results. Then the process is replicated T times. The results will be a forecast chart of each decision variable with T values. In other words, a simulation is run and the forecast or assumption statistics are used in the optimization model to find the optimal allocation of decision variables. Then, another simulation is run, generating different forecast statistics, and these new updated values are then optimized, and so forth. Hence, each of the final decision variables will have its own forecast chart, indicating the range of the optimal decision variables. For instance, instead of obtaining single-point estimates in the dynamic optimization procedure, you can now obtain a distribution of the decision variables, hence, a range of optimal values for each decision variable, also known as a stochastic optimization.

Finally, an efficient frontier optimization procedure applies the concepts of marginal increments and shadow pricing in optimization. That is, what would happen to the results of the optimization if one of the constraints were relaxed slightly? Say, for instance, the budget constraint is set at \$1 million. What would happen to the portfolio's outcome and optimal decisions if the constraint were now \$1.5 million or \$2 million, and so forth? This is the concept of the Markowitz efficient

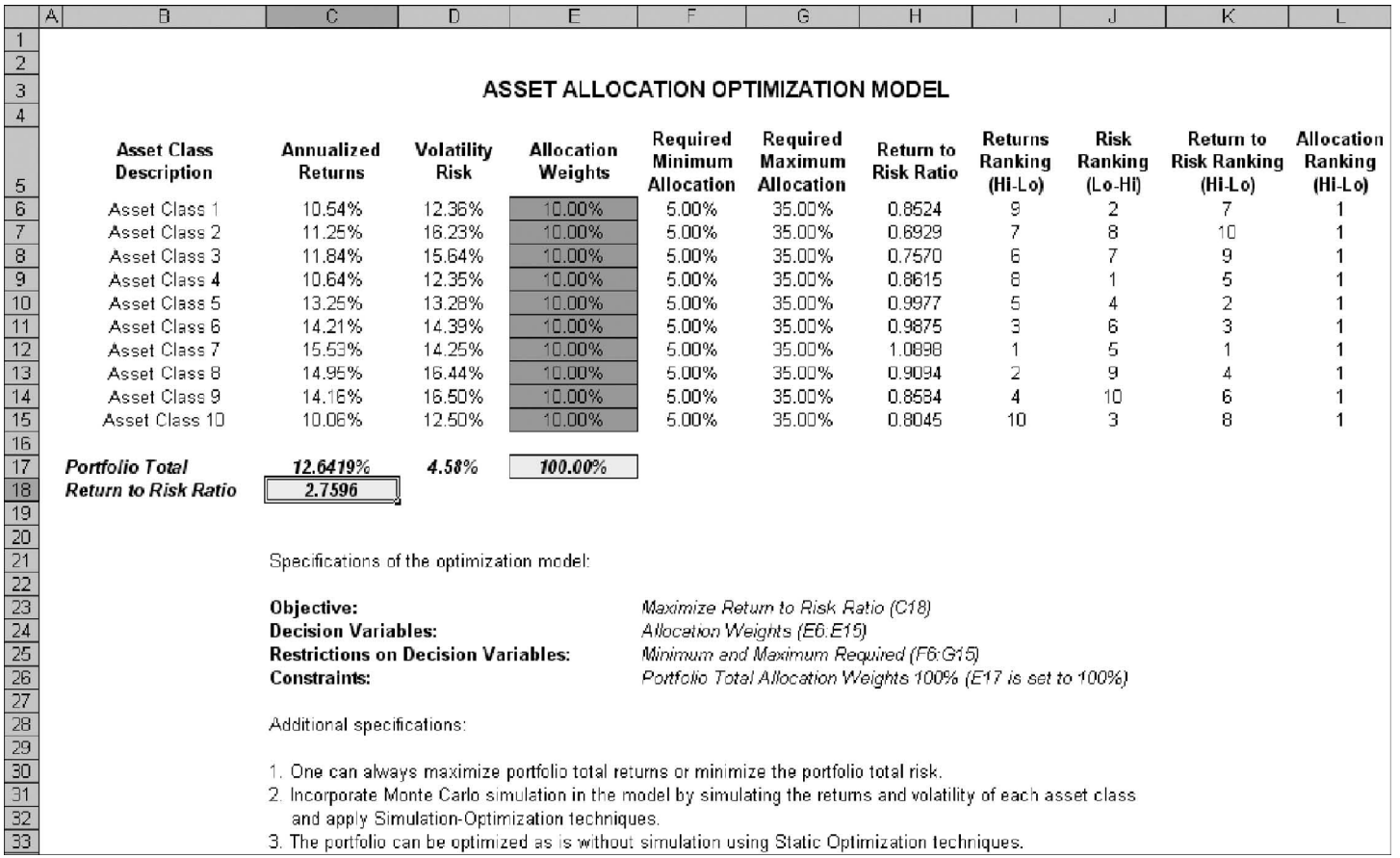
frontiers in investment finance, where if the portfolio standard deviation is allowed to increase slightly, what additional returns will the portfolio generate? This process is similar to the dynamic optimization process with the exception that *one* of the constraints is allowed to change, and with each change, the simulation and optimization process is run. This process is best applied manually using Risk Simulator. That is, run a dynamic or stochastic optimization, then rerun another optimization with a constraint, and repeat that procedure several times. This manual process is important, as by changing the constraint, the analyst can determine if the results are similar or different, and hence, whether it is worthy of any additional analysis, or can determine how far a marginal increase in the constraint should be to obtain a significant change in the objective and decision variables.

One item is worthy of consideration. There exist other software products that supposedly perform stochastic optimization but in fact they do not. For instance, after a simulation is run, then *one* iteration of the optimization process is generated, and then another simulation is run, then the *second* optimization iteration is generated and so forth; this is simply a waste of time and resources. That is, in optimization, the model is put through a rigorous set of algorithms, where multiple iterations (ranging from several to thousands of iterations) are required to obtain the optimal results. Hence, generating *one* iteration at a time is a waste of time and resources. The same portfolio can be solved using Risk Simulator in under a minute as compared to multiple hours using such a backward approach. Also, such a simulation-optimization approach will typically yield bad results, and is not a stochastic optimization approach. Be extremely careful of such methodologies when applying optimization to your models.

The following are two example optimization problems. One uses continuous decision variables, while the other uses discrete integer decision variables. In either model, you can apply discrete optimization, dynamic optimization, stochastic optimization, or even the efficient frontiers with shadow pricing. Any of these approaches can be used for these two examples. Therefore, for simplicity, only the model setup will be illustrated and it is up to the user to decide which optimization process to run. Also, the continuous model uses the nonlinear optimization approach (this is because the portfolio risk computed is a nonlinear function, and the objective is a nonlinear function of portfolio returns divided by portfolio risks), while the second example of an integer optimization is an example of a linear optimization model (its objective and all of its constraints are linear). Therefore, these two examples encapsulate all of the procedures aforementioned.

Example: Optimization with Continuous Decision Variables Figure I.54 illustrates the sample continuous optimization model. The example here uses the ***Continuous Optimization*** file accessed through **Risk Simulator | Examples.** In this example, there are 10 distinct asset classes (e.g., different types of mutual funds, stocks, or assets) where the idea is to most efficiently and effectively allocate the portfolio holdings such that the best bang for the buck is obtained. That is, to generate the best portfolio returns possible given the risks inherent in each asset class. In order to truly understand the concept of optimization, we will have to delve more deeply into this sample model in order to see how the optimization process can best be applied.

The model shows the 10 asset classes, and each asset class has its own set of annualized returns and annualized volatilities. These return and risk measures are



ASSET ALLOCATION OPTIMIZATION MODEL

Asset Class Description	Annualized Returns	Volatility Risk	Allocation Weights	Required Minimum Allocation	Required Maximum Allocation	Return to Risk Ratio	Returns Ranking (Hi-Lo)	Risk Ranking (Lo-Hi)	Return to Risk Ranking (Hi-Lo)	Allocation Ranking (Hi-Lo)
Asset Class 1	10.54%	12.36%	10.00%	5.00%	35.00%	0.8524	9	2	7	1
Asset Class 2	11.25%	16.23%	10.00%	5.00%	35.00%	0.6929	7	8	10	1
Asset Class 3	11.84%	15.64%	10.00%	5.00%	35.00%	0.7570	6	7	9	1
Asset Class 4	10.64%	12.35%	10.00%	5.00%	35.00%	0.8615	8	1	5	1
Asset Class 5	13.25%	13.28%	10.00%	5.00%	35.00%	0.9977	5	4	2	1
Asset Class 6	14.21%	14.39%	10.00%	5.00%	35.00%	0.9875	3	6	3	1
Asset Class 7	15.53%	14.25%	10.00%	5.00%	35.00%	1.0898	1	5	1	1
Asset Class 8	14.95%	16.44%	10.00%	5.00%	35.00%	0.9094	2	9	4	1
Asset Class 9	14.16%	16.50%	10.00%	5.00%	35.00%	0.8584	4	10	6	1
Asset Class 10	10.06%	12.50%	10.00%	5.00%	35.00%	0.8045	10	3	8	1
Portfolio Total	***12.6419%***	***4.58%***	***100.00%***							
Return to Risk Ratio	***2.7596***									

Specifications of the optimization model:

Objective: *Maximize Return to Risk Ratio (C18)*
Decision Variables: *Allocation Weights (E6:E15)*
Restrictions on Decision Variables: *Minimum and Maximum Required (F6:G15)*
Constraints: *Portfolio Total Allocation Weights 100% (E17 is set to 100%)*

Additional specifications:

1. One can always maximize portfolio total returns or minimize the portfolio total risk.
2. Incorporate Monte Carlo simulation in the model by simulating the returns and volatility of each asset class and apply Simulation-Optimization techniques.
3. The portfolio can be optimized as is without simulation using Static Optimization techniques.

FIGURE I.54 Continuous optimization model

annualized values such that they can be consistently compared across different asset classes. Returns are computed using the geometric average of the relative returns, while the risks are computed using the logarithmic relative stock returns approach. See the chapters on Volatility Models for details on computing the annualized volatility and annualized returns on a stock or asset class.

The Allocation Weights in column E holds the decision variables, which are the variables that need to be tweaked and tested such that the total weight is constrained at 100% (cell E17). Typically, to start the optimization, we will set these cells to a uniform value, where in this case, cells E6 to E15 are set at 10% each. In addition, each decision variable may have specific restrictions in its allowed range. In this example, the lower and upper allocations allowed are 5% and 35%, as seen in columns F and G. This means that each asset class may have its own allocation boundaries. Next, column H shows the return to risk ratio, which is simply the return percentage divided by the risk percentage, where the higher this value, the higher the bang for the buck. The remaining model shows the individual asset class rankings by returns, risk, return to risk ratio, and allocation. In other words, these rankings show at a glance which asset class has the lowest risk, the highest return, and so forth.

The portfolio's total returns in cell C17 is *SUMPRODUCT(C6:C15, E6:E15)*, that is, the sum of the allocation weights multiplied by the annualized returns for each asset class. As an example, with a portfolio of four assets, we have $R_P = \omega_A R_A + \omega_B R_B + \omega_C R_C + \omega_D R_D$, where R_P is the return on the portfolio, $R_{A,B,C,D}$ are the individual returns on the projects, and $\omega_{A,B,C,D}$ are the respective weights or capital allocation across each project.

In addition, the portfolio's diversified risk in cell D17 is computed by taking $\sigma_P = \sqrt{\sum_{i=1}^{i} \omega_i^2 \sigma_i^2 + \sum_{i=1}^{n} \sum_{j=1}^{m} 2\omega_i \omega_j \rho_{i,j} \sigma_i \sigma_j}$. Here, $\rho_{i,j}$ are the respective cross-correlations between the asset classes—hence, if the cross-correlations are negative, there are risk diversification effects, and the portfolio risk decreases. However, to simplify the computations here, we assume zero correlations among the asset classes through this portfolio risk computation, but assume the correlations when applying simulation on the returns as will be seen later. Therefore, instead of applying static correlations among these different asset returns, we apply the correlations in the simulation assumptions themselves, creating a more dynamic relationship among the simulated return values.

Finally, the return to risk ratio or Sharpe ratio is computed for the portfolio. This value is seen in cell C18, and represents the objective to be maximized in this optimization exercise. To summarize, we have the following specifications in this example model:

Objective:	Maximize Return to Risk Ratio (C18)
Decision Variables:	Allocation Weights (E6:E15)
Restrictions on Decision Variables:	Minimum and Maximum Required (F6:G15)
Constraints:	Total Allocation Weights Sum to 100% (E17)

The model has been preset to run the optimization (simply click on **Risk Simulator | Optimization | Run Optimization**) or alternatively, the following shows how to recreate the optimization model.

PROCEDURE

- Open the example file and start a new profile by clicking on **Risk Simulator | New Profile** and provide it a name.
- The first step in optimization is to set the decision variables. Select cell E6 and set the first decision variable (**Risk Simulator | Optimization | Set Decision**) and click on the link icon to select the name cell (B6), as well as the lower bound and upper bound values at cells F6 and G6. Then, using Risk Simulator copy, copy this cell E6 decision variable and paste the decision variable to the remaining cells in E7 to E15. Make sure to use Risk Simulator copy and paste, rather than Excel's copy and paste.
- The second step in optimization is to set the constraint. There is only one constraint here; that is, the total allocation in the portfolio must sum to 100%. So, click on **Risk Simulator | Optimization | Constraints**. . . and select **ADD** to add a new constraint. Then, select the cell E17 and make it equal (=) to 100%. Click **OK** when done.
- The final step in optimization is to set the objective function and start the optimization by selecting the objective cell C18, select **Risk Simulator | Optimization | Set Objective** and **Risk Simulator | Optimization | Run Optimization** and choosing the optimization of choice (Static Optimization, Dynamic Optimization, or Stochastic Optimization). To get started, select **Static Optimization.** Check to make sure the objective cell is set for C18 and select **Maximize**. You can now review the decision variables and constraints if required, or click **OK** to run the static optimization.
- Once the optimization is complete, you may select **Revert** to revert back to the original values of the decision variables as well as the objective, or select **Replace** to apply the optimized decision variables. Typically, **Replace** is chosen after the optimization is done.

Figure I.55 shows the screen shots of these procedural steps. You can add simulation assumptions on the model's returns and risk (columns C and D) and apply the dynamic optimization and stochastic optimization for additional practice.

Results Interpretation The optimization's final results are shown in Figure I.56, where the optimal allocation of assets for the portfolio is seen in cells E6:E15. That is, given the restrictions of each asset fluctuating between 5% and 35%, and where the sum of the allocation must equal 100%, the allocation that maximizes the return to risk ratio is seen in Figure I.56.

A few important points have to be noted when reviewing the results and optimization procedures performed thus far:

- The correct way to run the optimization is to maximize the bang for the buck or returns to risk Sharpe ratio as we have done.
- If instead we maximized the total portfolio returns, the optimal allocation result is trivial and does not require optimization to obtain. That is, simply allocate 5% (the minimum allowed) to the lowest eight assets, 35% (the maximum allowed) to the highest-returning asset, and the remaining (25%) to the second-best returns asset. Optimization is not required. However, when allocating the

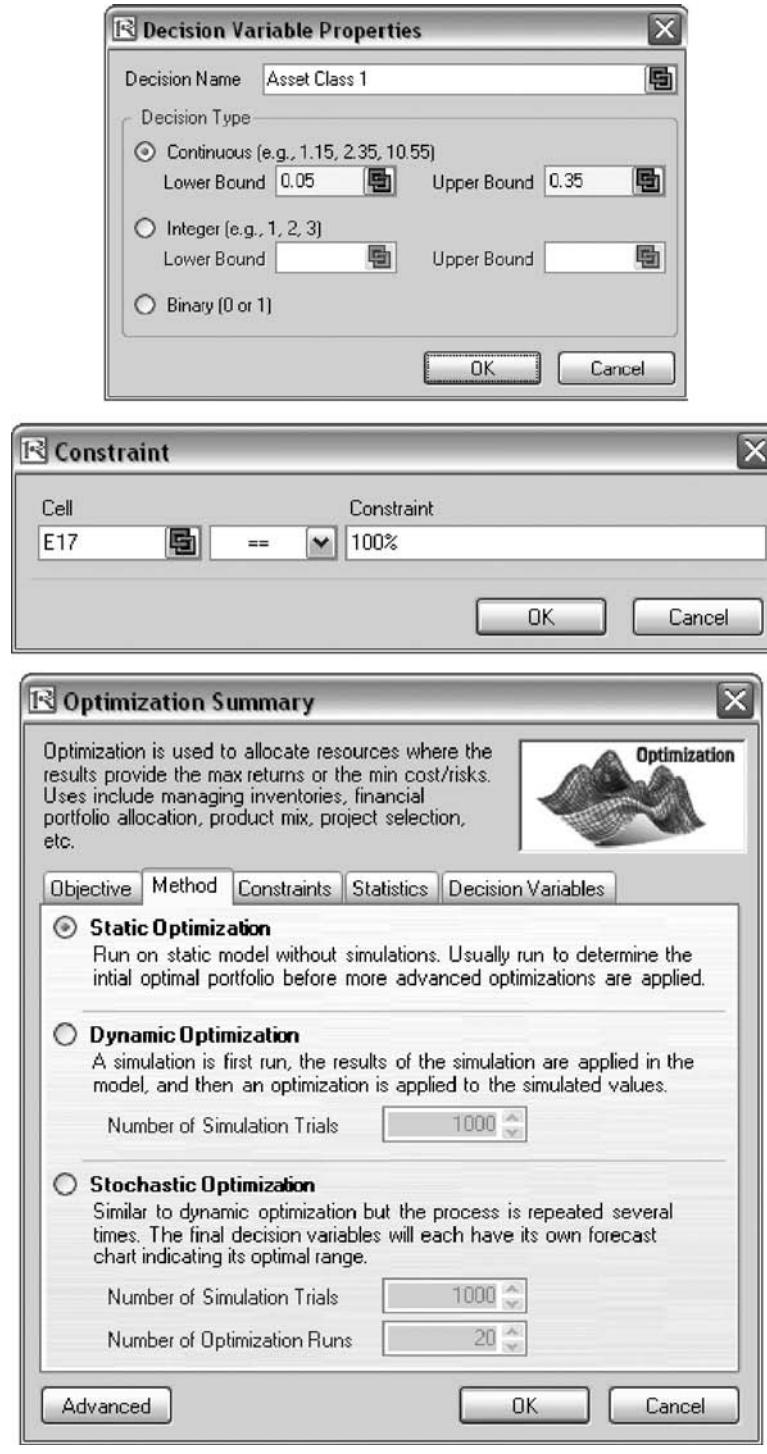


FIGURE I.55 Running Continuous Optimization in Risk Simulator

ASSET ALLOCATION OPTIMIZATION MODEL

Asset Class Description	Annualized Returns	Volatility Risk	Allocation Weights	Required Minimum Allocation	Required Maximum Allocation	Return to Risk Ratio	Returns Ranking (Hi-Lo)	Risk Ranking (Lo-Hi)	Return to Risk Ranking (Hi-Lo)	Allocation Ranking (Hi-Lo)
Asset Class 1	10.54%	12.36%	11.09%	5.00%	35.00%	0.8524	9	2	7	4
Asset Class 2	11.25%	16.23%	6.87%	5.00%	35.00%	0.6929	7	8	10	10
Asset Class 3	11.84%	15.64%	7.78%	5.00%	35.00%	0.7570	6	7	9	9
Asset Class 4	10.64%	12.35%	11.22%	5.00%	35.00%	0.8615	8	1	5	3
Asset Class 5	13.25%	13.28%	12.08%	5.00%	35.00%	0.9977	5	4	2	2
Asset Class 6	14.21%	14.39%	11.04%	5.00%	35.00%	0.9875	3	6	3	5
Asset Class 7	15.53%	14.25%	12.30%	5.00%	35.00%	1.0898	1	5	1	1
Asset Class 8	14.95%	16.44%	8.90%	5.00%	35.00%	0.9094	2	9	4	7
Asset Class 9	14.16%	16.50%	8.37%	5.00%	35.00%	0.8584	4	10	6	8
Asset Class 10	10.06%	12.50%	10.35%	5.00%	35.00%	0.8045	10	3	8	6
Portfolio Total	*12.6920%*	*4.52%*	*100.00%*							
Return to Risk Ratio	*2.8091*									

FIGURE I.56 Continuous optimization results

TABLE I.3 Optimization Results

Objective	Portfolio Returns	Portfolio Risk	Portfolio Returns to Risk Ratio
Maximize Returns to Risk Ratio	12.69%	4.52%	2.8091
Maximize Returns	13.97%	6.77%	2.0636
Minimize Risk	12.38%	4.46%	2.7754

portfolio this way, the risk is a lot higher as compared to when maximizing the returns to risk ratio, although the portfolio returns by themselves are higher.

- In contrast, one can minimize the total portfolio risk, but the returns will now be less.

Table I.3 illustrates the results from the three different objectives being optimized.

From the table, the best approach is to maximize the returns to risk ratio; that is, for the same amount of risk, this allocation provides the highest amount of return. Conversely, for the same amount of return, this allocation provides the lowest amount of risk possible. This approach of bang for the buck or returns to risk ratio is the cornerstone of the Markowitz efficient frontier in modern portfolio theory. That is, if we constrained the total portfolio risk levels and successively increased it over time we will obtain several efficient portfolio allocations for different risk characteristics. Thus, different efficient portfolio allocations can be obtained for different individuals with different risk preferences.

OPTIMIZATION WITH DISCRETE INTEGER VARIABLES

Sometimes, the decision variables are not continuous but discrete integers (e.g., 1,2,3) or binary (e.g., 0 and 1). That is, we can use such optimization as on-off switches or go/no-go decisions. Figure I.57 illustrates a project selection model where there are 12 projects listed. The example here uses the **Discrete Optimization** file found either on the start menu at **Start | Real Options Valuation | Risk Simulator | Examples** or accessed directly through **Risk Simulator | Example Models.** Each project, like before, has its own returns (ENPV and NPV for expanded net present value and net present value—the ENPV is simply the NPV plus any strategic real options values), costs of implementation, risks, and so forth. If required, this model can be modified to include required full-time equivalences (FTE) and other resources of various functions, and additional constraints can be set on these additional resources. The inputs into this model are typically linked from other spreadsheet models. For instance, each project will have its own discounted cash flow or returns on investment model. The application here is to maximize the portfolio's Sharpe ratio subject to some budget allocation. Many other versions of this model can be created—for instance, maximizing the portfolio returns, minimizing the risks, or adding additional constraints where the total number of projects chosen cannot exceed 6. All of these items can be run using this existing model.

	Credit Line	ENPV	Cost	Risk $	Risk %	Return to Risk Ratio	Profitability Index	Selection
4	Project 1	$458.00	$1,732.44	$54.96	12.00%	8.33	1.26	1.0000
5	Project 2	$1,954.00	$859.00	$1,914.92	98.00%	1.02	3.27	1.0000
6	Project 3	$1,599.00	$1,845.00	$1,551.03	97.00%	1.03	1.87	1.0000
7	Project 4	$2,251.00	$1,645.00	$1,012.95	45.00%	2.22	2.37	1.0000
8	Project 5	$849.00	$458.00	$925.41	109.00%	0.92	2.85	1.0000
9	Project 6	$758.00	$52.00	$560.92	74.00%	1.35	15.58	1.0000
10	Project 7	$2,845.00	$758.00	$5,633.10	198.00%	0.51	4.75	1.0000
11	Project 8	$1,235.00	$115.00	$926.25	75.00%	1.33	11.74	1.0000
12	Project 9	$1,945.00	$125.00	$2,100.60	108.00%	0.93	16.56	1.0000
13	Project 10	$2,250.00	$458.00	$1,912.50	85.00%	1.18	5.91	1.0000
14	Project 11	$549.00	$45.00	$263.52	48.00%	2.08	13.20	1.0000
15	Project 12	$525.00	$105.00	$309.75	59.00%	1.69	6.00	1.0000
17	Total	$17,218.00	$8,197.44	$7,007	40.70%			12.00
18	Goal:	MAX	< =$5000					<=6
19	Sharpe Ratio	2.4573						

ENPV is the expected NPV of each credit line or project, while Cost can be the total cost of administration as well as required capital holdings to cover the credit line, and Risk is the Coefficient of Variation of the credit line's ENPV.

FIGURE 1.57 Discrete integer optimization model

The example model had been preset and is ready to run (**Risk Simulator | Change Profile** and select the optimization profile and then click on **Risk Simulator | Optimization | Run Optimization**) or follow the procedure below to recreate the optimization model form scratch).

PROCEDURE

- Open the example file and start a new profile by clicking on **Risk Simulator | New Profile** and provide it a name.
- The first step in optimization is to set up the decision variables. Set the first decision variable by selecting cell J4, and select **Risk Simulator | Optimization | Set Decision**, click on the link icon to select the name cell (B4), and select the **Binary** variable. Then, using Risk Simulator copy, copy this cell J4 decision variable and paste the decision variable to the remaining cells in J5 to J15. This is the best method if you have only several decision variables and you can name each decision variable with a unique name for identification later. Make sure to use the Risk Simulator copy and paste, rather than Excel copy and paste functions.
- The second step in optimization is to set the constraints. There are two constraints here; that is, the total budget allocation in the portfolio must be less than $5,000 and the total number of projects must not exceed 6. So, click on **Risk Simulator | Optimization | Constraints**... and select **ADD** to add a new constraint. Then, select the cell D17 and make it less than or equal ($\leq$) to 5000. Repeat by setting cell J17 $\leq$ 6.

- The final step in optimization is to set the objective function and start the optimization by selecting cell C19 and selecting **Risk Simulator | Optimization | Set Objective**; then, run the optimization using **Risk Simulator | Optimization | Run Optimization** and choosing the optimization of choice (Static Optimization, Dynamic Optimization, or Stochastic Optimization). To get started, select **Static Optimization.** Check to make sure that the objective cell is either the Sharpe ratio or portfolio returns to risk ratio and select **Maximize**. You can now review the decision variables and constraints if required, or click **OK** to run the static optimization.

Figure I.58 shows the screen shots of these procedural steps. You can add simulation assumptions on the model's ENPV and risk (columns C and D) and apply the dynamic optimization and stochastic optimization for additional practice.

Results Interpretation Figure I.59 shows a sample optimal selection of projects that maximizes the Sharpe ratio. In contrast, one can always maximize total revenues, but as before, this is a trivial process and simply involves choosing the highest returning project and going down the list until you run out of money or exceed the budget constraint. Doing so will yield theoretically undesirable projects as the highest-yielding projects typically hold higher risks. Now, if desired, you can replicate the optimization using a stochastic or dynamic optimization by adding in assumptions in the ENPV and Risk values.

For additional hands-on examples of optimization in action, see the various chapters on optimization models.

FORECASTING

Forecasting is the act of predicting the future, whether it is based on historical data or on speculation about the future when no history exists. When historical data exist, a quantitative or statistical approach is best, but if no historical data exist, then potentially a qualitative or judgmental approach is usually the only recourse.

Different Types of Forecasting Techniques

Generally, forecasting can be divided into quantitative and qualitative techniques. Qualitative forecasting is used when little to no reliable historical, contemporaneous, or comparable data exists. Several qualitative methods exist, such as the Delphi or expert opinion approach (a consensus-building forecast by field experts, marketing experts, or internal staff members); management assumptions (target growth rates set by senior management); as well as market research or external data or polling and surveys (data obtained through third-party sources, industry and sector indexes, or from active market research). These estimates can be either single-point estimates (an average consensus) or a set of forecast values (a distribution of forecasts). The latter can be entered into Risk Simulator as a custom distribution, and the resulting forecasts can be simulated. That is, a nonparametric simulation can be run using the estimated data points themselves as the distribution.

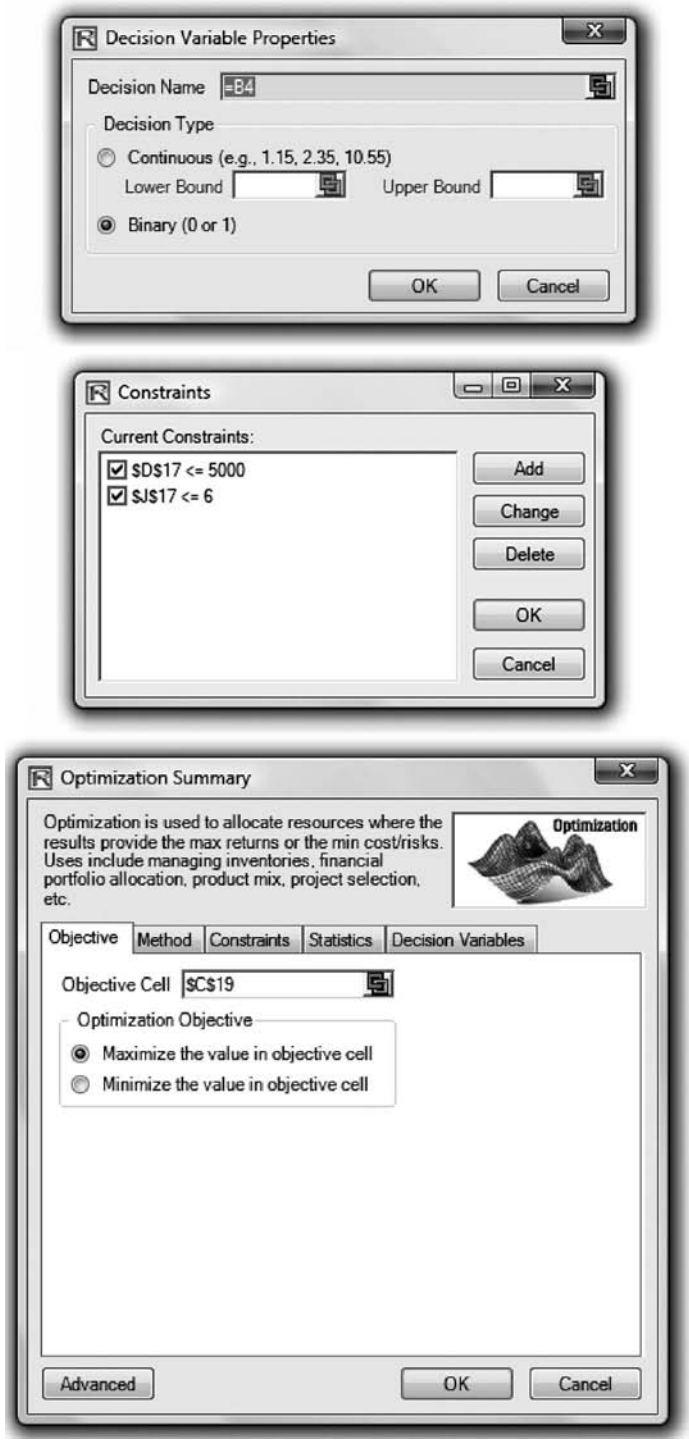


FIGURE I.58 Running discrete integer optimization

Credit Line	ENPV	Cost	Risk $	Risk %	Return to Risk Ratio	Profitability Index	Selection
Project 1	$458.00	$1,732.44	$54.96	12.00%	8.33	1.26	1.0000
Project 2	$1,954.00	$859.00	$1,914.92	98.00%	1.02	3.27	0.0000
Project 3	$1,599.00	$1,845.00	$1,551.03	97.00%	1.03	1.87	0.0000
Project 4	$2,251.00	$1,645.00	$1,012.95	45.00%	2.22	2.37	1.0000
Project 5	$849.00	$458.00	$925.41	109.00%	0.92	2.85	0.0000
Project 6	$758.00	$52.00	$560.92	74.00%	1.35	15.58	1.0000
Project 7	$2,845.00	$758.00	$5,633.10	198.00%	0.51	4.75	0.0000
Project 8	$1,235.00	$115.00	$926.25	75.00%	1.33	11.74	1.0000
Project 9	$1,945.00	$125.00	$2,100.60	108.00%	0.93	16.56	0.0000
Project 10	$2,250.00	$458.00	$1,912.50	85.00%	1.18	5.91	0.0000
Project 11	$549.00	$45.00	$263.52	48.00%	2.08	13.20	1.0000
Project 12	$525.00	$105.00	$309.75	59.00%	1.69	6.00	1.0000
Total	$5,776.00	$3,694.44	$1,539	26.64%			6.00
Goal:	MAX	<=$5000					<=6
Sharpe Ratio	3.7543						

ENPV is the expected NPV of each credit line or project, while Cost can be the total cost of administration as well as required capital holdings to cover the credit line, and Risk is the Coefficient of Variation of the credit line's ENPV.

FIGURE 1.59 Optimal selection of projects that maximizes the Sharpe ratio

On the quantitative side of forecasting, the available data or data that need to be forecasted can be divided into time-series (values that have a time element to them, such as revenues of different years, inflation rates, interest rates, market share, and so forth, or failure rates); cross-sectional (values that are time-independent, such as the grade point average of sophomore students across the nation in a particular year, given each student's levels of SAT scores, IQ, and number of alcoholic beverages consumed per week); or mixed panel (mixture between time-series and panel data, such as predicting sales over the next 10 years given budgeted marketing expenses and market share projections; this means that the sales data is time-series, but exogenous variables such as marketing expenses and market share exist to help to model the forecast predictions).

The Risk Simulator software provides the user several forecasting methodologies:

- Auto-ARIMA
- Basic Econometrics
- Box-Jenkins ARIMA
- Custom Distributions
- J-S Curves
- GARCH
- Markov Chains
- Maximum Likelihood
- Multivariate Regression
- Nonlinear Extrapolation
- Spline Curves
- Stochastic Process Forecasting
- Time-Series Analysis

In general, to create forecasts, several quick steps are required:

- Start Excel and enter in or open your existing historical data.
- Select the data, and click on **Risk Simulator** and select **Forecasting**.

- Select the relevant forecasting application (Auto-ARIMA, Basic Econometrics, Box-Jenkins ARIMA, J-S Curves, GARCH, Markov Chains, Maximum Likelihood, Multivariate Regression, Nonlinear Extrapolation, Spline Curves, Stochastic Process Forecasting, or Time-Series Analysis), and enter the relevant inputs.

The following provides a quick review of each methodology and several quick getting-started examples in using the software. More detailed descriptions and example models of each of these techniques are found throughout this book.

Auto-ARIMA

- Autoregressive integrated moving average (ARIMA) is an advanced econometric modeling technique. ARIMA looks at historical time-series data and performs back-fitting optimization routines to account for historical autocorrelation (the relationship of one value versus another in time), the stability of the data to correct for the nonstationary characteristics of the data, and this predictive model learns over time by correcting its forecasting errors. Advanced knowledge in econometrics is typically required to build good predictive models using this approach.
- The Auto-ARIMA module automates some of the traditional ARIMA modeling by automatically testing multiple permutations of model specifications and returns the best-fitting model. Running the Auto-ARIMA is similar to regular ARIMA forecasts, the difference being the P, D, Q inputs are no longer required and different combinations of these inputs are automatically run and compared.
- See Chapter 90, "Forecasting—Time-Series ARIMA," for more technical details on running and interpreting an ARIMA model. This approach can only be used to forecast time-series data and can include other independent variables in its forecasts.

Basic Econometrics

- Econometrics refers to a branch of business analytics, modeling, and forecasting techniques for modeling the behavior or forecasting certain business or economic variables.
- Running the Basic Econometrics models are similar to doing regular regression analysis except that the dependent and independent variables are allowed to be modified before a regression is run.
- See Chapter 87, "Forecasting—Multiple Regression," for details on running regression models. This approach can be used to model the relationship or forecast time-series, cross-sectional, as well as mixed data sets.

Box-Jenkins ARIMA

- A summary of this methodology is provided earlier, in the Auto-ARIMA section.
- See Chapter 90, "Forecasting—Time-Series ARIMA," for details on running an ARIMA model. This approach can only be used to forecast time-series data and can include other independent variables in its forecasts.

Custom Distributions

- Using Risk Simulator, expert opinions can be collected and a customized distribution can be generated. This forecasting technique comes in handy when the data set is small or the goodness-of-fit is bad when applied to a distributional fitting routine.
- See Chapter 132, "Risk Hedging—Foreign Exchange Cash Flow Model," for details on creating a custom distribution (in the chapter, a custom distribution is created to forecast foreign exchange rates). This approach can be used to forecast time-series, cross-sectional, or mixed data sets.

J-S Curves

- The J-curve or exponential growth curve is where the growth of the next period depends on the current period's level and the increase is exponential. This means that over time, the values will increase significantly from one period to another. This model is typically used in forecasting biological growth and chemical reactions over time.
- The S-curve or logistic growth curve starts off like a J-curve, with exponential growth rates. Over time, the environment becomes saturated (e.g., market saturation, competition, overcrowding); the growth slows; and the forecast value eventually ends up at a saturation or maximum level. This model is typically used in forecasting market share or sales growth of a new product from market introduction until maturity and decline.
- See Chapter 82, "Forecasting—Exponential J-Growth Curves," for details on running the J-curve model, and Chapter 85, "Forecasting—Logistic S-Growth Curves," for running the S-curve model. These approaches are used to forecast time-series data.

GARCH

- The generalized autoregressive conditional heteroskedasticity (GARCH) model is used to model historical and forecast future volatility levels of a marketable security (e.g., stock prices, commodity prices, oil prices, and so forth). The data set has to be a time-series of raw price levels. GARCH will first convert the prices into relative returns and then run an internal optimization to fit the historical data to a mean-reverting volatility term structure, while assuming that the volatility is heteroskedastic in nature (changes over time according to some econometric characteristics).
- See Chapter 166, "Volatility—Volatility Computations (Log Returns, Log Assets, Implied Volatility, Management Assumptions, EWMA, GARCH)," for details on the GARCH model. This approach is used for forecasting the time-series of volatility of a marketable security. There must be a lot of data available and the data points must all be positive.

Markov Chains

- A Markov chain exists when the probability of a future state depends on a previous state and when linked together forms a chain that reverts to a long-run steady state level. This approach is typically used to forecast the market share of two competitors. The required inputs are the starting probability of a customer

in the first store (the first state) will return to the same store in the next period, versus the probability of switching to a competitor's store in the next state.
- See Chapter 86, "Forecasting—Markov Chains and Market Share," for details. This method is used to forecast a time-series of probabilistic states and the long-run steady-state condition.

Maximum Likelihood

- Maximum likelihood estimation (MLE) is used to forecast the probability of something occurring given some independent variables. For instance, MLE is used to predict if a credit line or debt will default given the obligor's characteristics (30 years old, single, salary of $100,000 per year, and has a total credit card debt of $10,000); or the probability a patient will have lung cancer if the person is a male, between the ages of 50 and 60, smokes five packs of cigarettes per month, and so forth.
- See Chapter 118. "Probability of Default—Empirical Model," for details on running this MLE model. The data set are typically cross-sectional and the dependent variable has to be binary (with values of 0 or 1).

Multivariate Regression

- Multivariate regression is used to model the relationship structure and characteristics of a certain dependent variable as it depends on other independent exogenous variables. Using the modeled relationship, we can forecast the future values of the dependent variable. The accuracy and goodness of fit for this model can also be determined. Linear and nonlinear models can be fitted in regression analysis.
- See Chapter 87, "Forecasting—Multiple Regression," for details on running regression models. This methodology can be used to model and forecast time-series data, cross-sectional data, or mixed data.

Nonlinear Extrapolation

- The underlying structure of the data to be forecasted is assumed to be nonlinear over time. For instance, a data set such as 1, 4, 9, 16, 25 is considered to be nonlinear (these data points are from a squared function).
- See Chapter 88, "Forecasting—Nonlinear Extrapolation and Forecasting," for details on nonlinear extrapolation forecasts. This methodology is typically applied to forecast time-series data. Sometimes, cross-sectional data can be applied if there is a nonlinear relationship between data points arranged from small to large values.

Spline Curves

- Sometimes there are missing values in a time-series data set. For instance, interest rates for years 1 to 3 may exist, followed by years 5 to 8, and then year 10. Spline curves can be used to interpolate the missing years' interest rate values based on the data that exist. Spline curves can also be used to forecast or extrapolate values of future time periods beyond the time period of available data. The data can be linear or nonlinear.

- See Chapter 172, "Yield Curve—U.S. Treasury Risk-Free Rates and Cubic Spline Curves," for details. This methodology is used to back-fit and forecast time-series data only.

Stochastic Process Forecasting

- Sometimes variables are stochastic and cannot be readily predicted using traditional means. These variables are said to be stochastic. Nonetheless, most financial, economic and naturally occurring phenomena (e.g., motion of molecules through the air) follow a known mathematical law or relationship. Although the resulting values are uncertain, the underlying mathematical structure is known and can be simulated using Monte Carlo risk-simulation.
- See Chapter 89, "Forecasting—Stochastic Processes," for details on stochastic process forecasting where we forecast using random walk, Brownian motion, mean-reverting, and jump-diffusion processes.

Time-Series Analysis

- In well-behaved time-series data (typical examples include sales revenues and cost structures of large corporations), the values tend to have up to three elements, a base value, trend, and seasonality. Time-series analysis uses these historical data and decomposes them into these three elements, and recomposes them into future forecasts. In other words, this forecasting method, like some of the others discribed, first performs a back-fitting (backcast) of historical data before it provides estimates of future values (forecast).
- See Chapter 91, "Forecasting—Time-Series Analysis," for details on time-series decomposition models. This methodology is applicable only to time-series data.

1. Analytics—Central Limit Theorem

File Name: Analytics—Central Limit Theorem

Location: Modeling Toolkit | Analytics | Central Limit Theorem

Brief Description: Illustrating the concept of Central Limit Theorem and Law of Large Numbers using Risk Simulator's set assumptions functionality, where many distributions, at the limit, are shown to approach normality

Requirements: Modeling Toolkit, Risk Simulator

This example shows how the Central Limit Theorem works by using Risk Simulator and without the applications of any mathematical derivations. Specifically, we look at how the *normal* distribution sometimes can be used to approximate other distributions and how some distributions can be made to be highly flexible, as in the case of the beta distribution.

The Central Limit Theorem contains a set of weak-convergence results in probability theory. Intuitively, they all express the fact that any sum of many independent and identically distributed random variables will tend to be distributed according to a particular attractor distribution. The most important and famous result is called the Central Limit Theorem, which states that if the sum of the variables has a finite variance, then it will be approximately normally distributed. As many real processes yield distributions with finite variance, this theorem explains the ubiquity of the normal distribution. Also, the distribution of an average tends to be normal, even when the distribution from which the average is computed is decidedly not normal.

DISCRETE UNIFORM DISTRIBUTION

In this model, we look at various distributions and see that over a large sample size and various parameters, they approach normality. We start off with a highly unlikely candidate, the *discrete uniform* distribution. The discrete uniform distribution is also known as the equally likely outcomes distribution. Where the distribution has a set of N elements, and each element has the same probability (Figure 1.1). This distribution is related to the *uniform* distribution but its elements are discrete instead of continuous. The input requirement is such that minimum < maximum and both values must be integers. An example would be tossing a single die, with 6 sides. The probability of hitting 1, 2, 3, 4, 5, or 6 is exactly the same: 1/6. So, how can a distribution like this be converted into a normal distribution?

The idea lies in the combination of multiple distributions. Suppose you now take a pair of dice and toss them. You would have 36 possible outcomes; that is, the first

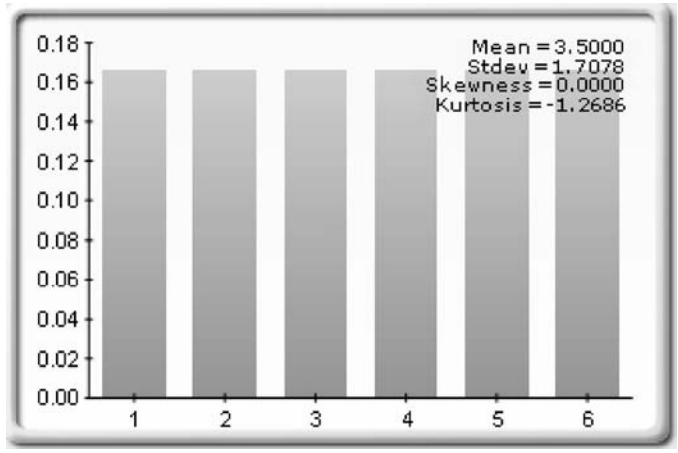


FIGURE 1.1 Tossing a single die and the discrete uniform distribution (values between 1 and 6)

	1	2	3	4	5	6
1	1,1	2,1	3,1	4,1	5,1	6,1
2	1,2	2,2	3,2	4,2	5,2	6,2
3	1,3	2,3	3,3	4,3	5,3	6,3
4	1,4	2,4	3,4	4,4	5,4	6,4
5	1,5	2,5	3,5	4,5	5,5	6,5
6	1,6	2,6	3,6	4,6	5,6	6,6

FIGURE 1.2 Tossing two dice (36 possible outcomes)

single die can be 1 and the second die can be 1, or perhaps 1–2, or 1–3, and so forth, until 6–6, with 36 outcomes as in Figure 1.2.

Now, summing up the two dice, you get an interesting set of results (Figure 1.3).

If you then plotted out these sums, you get an approximation of a normal distribution (Figure 1.4).

In fact, if you threw 12 dice together and added up their values, and repeated the process many times, you get an extremely close discrete normal distribution. If

	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

FIGURE 1.3
Summation of two dice

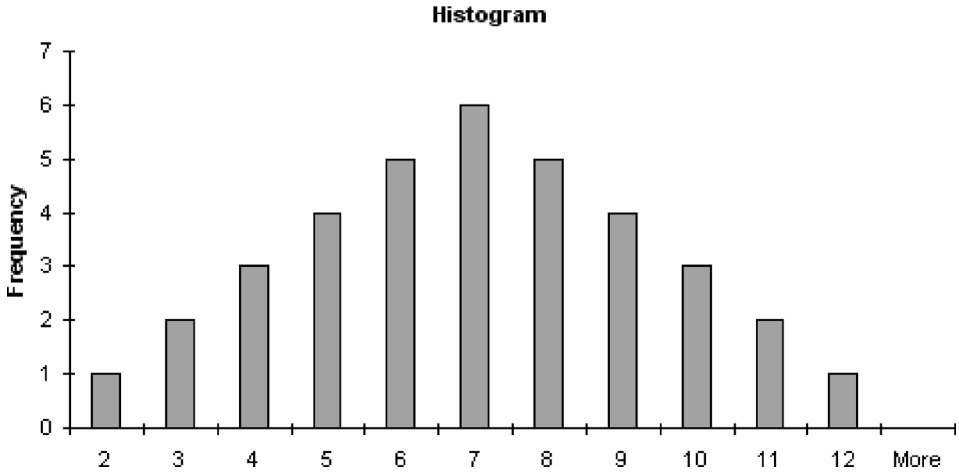


FIGURE 1.4 Approximation to a normal distribution

you add 12 continuous uniform distributions, where the results can, say, take on any continuous value between 1 and 6, you obtain a perfectly normal distribution.

POISSON, BINOMIAL, AND HYPERGEOMETRIC DISTRIBUTIONS

Continuing with the examples, we show that for higher values of the distributional parameters (where many trials exist), these three distributions also tend to normality. For instance, in the *Other Discrete* worksheet in the model, notice that as the number of trials (N) in a *binomial* distribution increases, the distribution tends to normal. Even with a small probability (P) value, as the number of trials N increases, normality again reigns (Figure 1.5). In fact, as $N \times P$ exceeds about 30, you can use the normal distribution to approximate the binomial distribution. Also, this is important, as at high N values, it is very difficult to compute the exact binomial distribution value, and the normal distribution is a lot easier to use.

We can test this approximation by using the *Distributional Analysis* tool (**Start | Programs | Real Options Valuation | Risk Simulator | Distribution Analysis**). As an example, we test a binomial distribution with $N = 5000$ and $P = 0.50$. We then compute the mean of the distribution, $NP = 2500$ and the standard deviation of the binomial distribution, $\sqrt{\mathrm{NP}(1-\mathrm{P})} = \sqrt{5000(0.5)(1-0.5)} = 35.3553$. We then enter these values in the normal distribution and look at the *Cumulative Distribution Function* (CDF) of some random range. Sure enough, the probabilities we obtain are close although not precisely the same (Figure 1.6).

The normal distribution does in fact approximate the binomial distribution when $N \times P$ is large (compare the results in Figures 1.6 and 1.7).

The examples also examine the *hypergeometric* and *Poisson* distributions. A similar phenomenon occurs. When the input parameters are large, they revert to

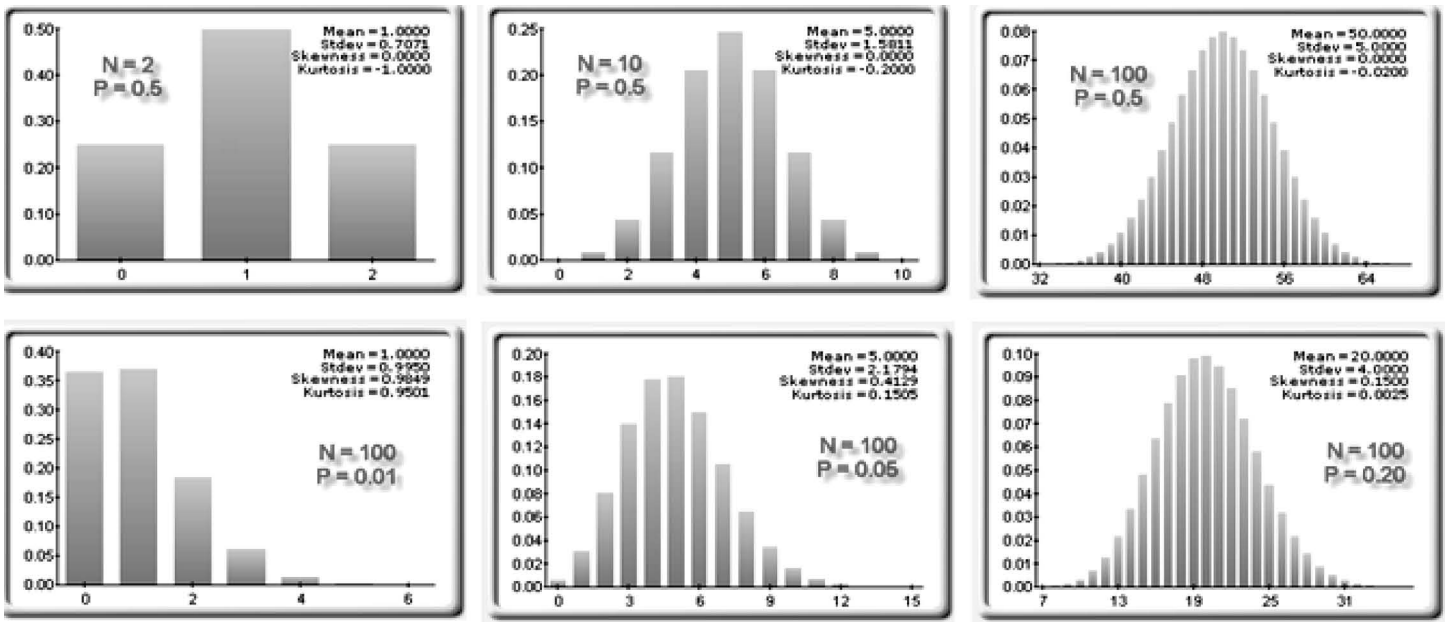


FIGURE 1.5 Different faces of a binomial distribution

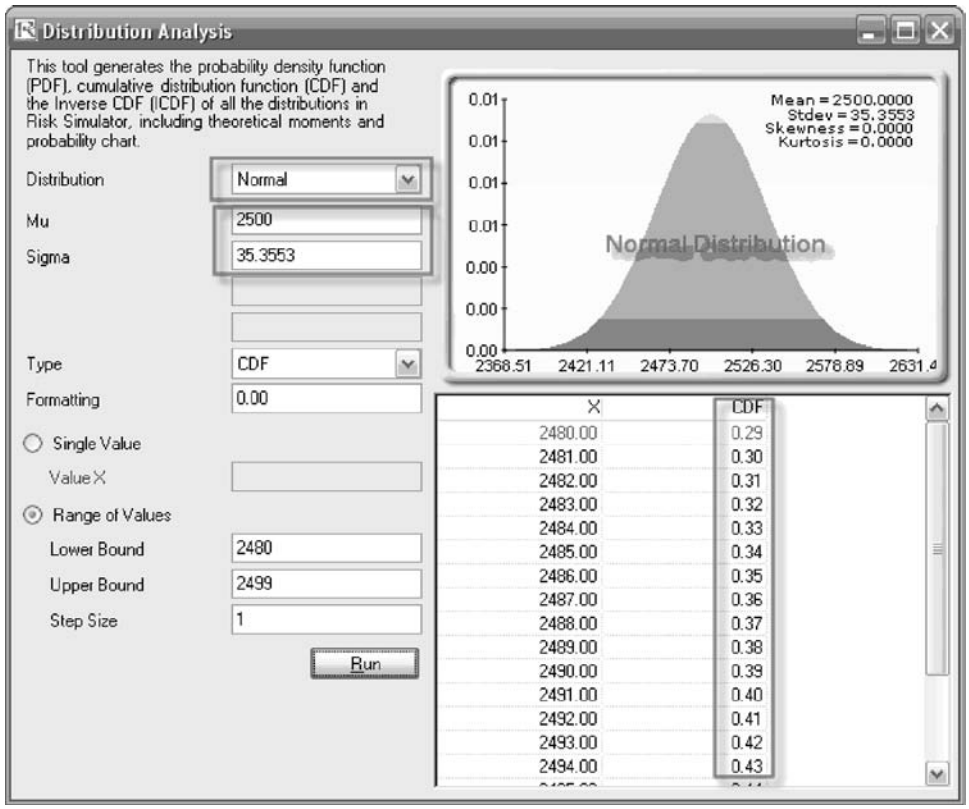


FIGURE 1.6 Normal approximation of the binomial

the normal approximation. In fact, the normal distribution also can be used to approximate the Poisson and hypergeometric distributions. Clearly there will be slight differences in value as the normal is a continuous distribution whereas the binomial, Poisson, and hypergeometric are discrete distributions. Therefore, slight variations will obviously exist.

BETA DISTRIBUTION

Finally, the *Beta* worksheet illustrates an interesting distribution, the beta distribution. Beta is a highly flexible and malleable distribution and can be made to approximate multiple distributions. If the two input parameters, *alpha* and *beta*, are equal, the distribution is symmetrical. If either parameter is 1 while the other parameter is greater than 1, the distribution is triangular or J-shaped. If alpha is less than beta, the distribution is said to be positively skewed (most of the values are near the minimum value). If alpha is greater than beta, the distribution is negatively skewed (most of the values are near the maximum value).

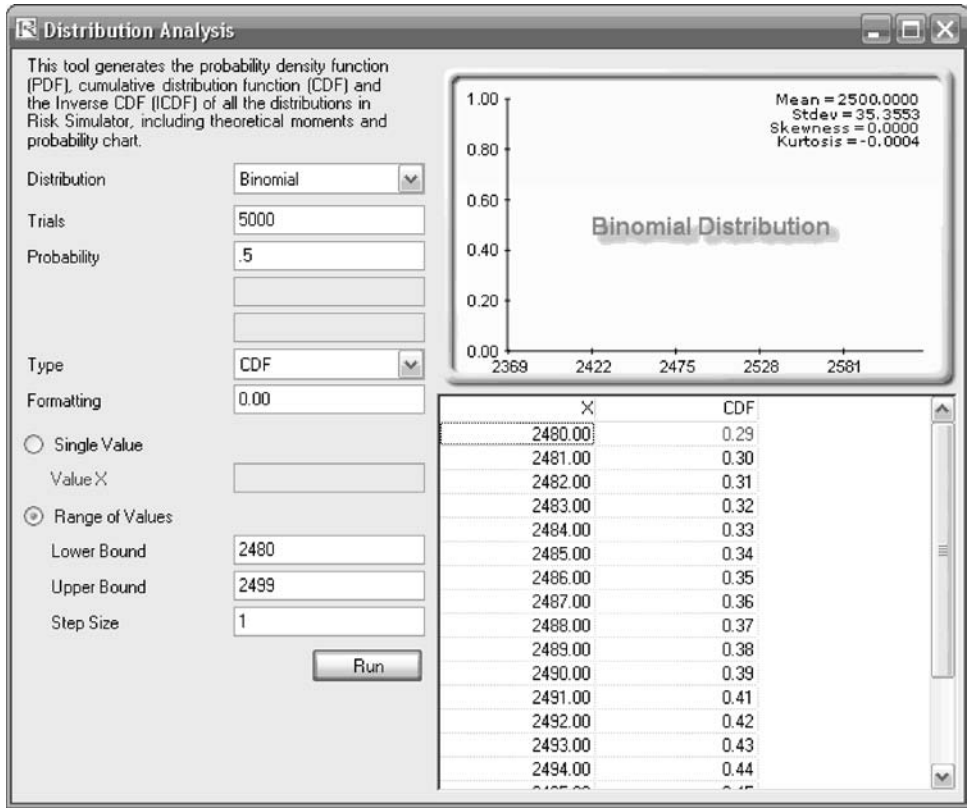


FIGURE 1.7 Binomial approximation of the normal

2. Analytics—Central Limit Theorem—Winning Lottery Numbers

File Name: Analytics—Central Limit Theorem (Winning Lottery Numbers)

Location: Modeling Toolkit | Analytics | Central Limit Theorem—Winning Lottery Numbers

Brief Description: Applying distributional fitting on past winning lottery numbers and to illustrate the Central Limit Theorem and Law of Large Numbers

Requirements: Modeling Toolkit, Risk Simulator

This fun model is used to illustrate the behavior of seemingly random events. For the best results, first review the Central Limit Theorem model in Chapter 1 before